Nr. 82
1. December 2021

universität innsbruck
Institut für Mathematik

Preprint-Series: Department of Mathematics - Applied Mathematics

A Variational View on Statistical Multiscale Estimation

M. Haltmeier, H. Li, A. Munk

AppliedMathematics

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

A Variational View on Statistical Multiscale Estimation

Markus Haltmeier

Department of Mathematics, University of Innsbruck
Technikestraße 13, A-6020 Innsbruck, Austria
Email: markus.haltmeier@uibk.ac.at

Housen Li, and Axel Munk

Institute of Mathematical Stochastics, University of Göttingen
Goldschmidtstr. 7, 37077 Göttingen, Germany
Email: housen.li@mathematik.uni-goettingen.de, munk@math.uni-goettingen.de

December 1, 2021

Abstract

We present a unifying view on various statistical estimation techniques including penalization, variational and thresholding methods. These estimators will be analyzed in the context of statistical linear inverse problems including nonparametric and change point regression, and high dimensional linear models as examples. Our approach reveals many seemingly unrelated estimation schemes as special instances of a general class of variational multiscale estimators, named MIND (MultIscale Nemirovskii–Dantzig). These estimators result from minimizing certain regularization functionals under convex constraints that can be seen as multiple statistical tests for local hypotheses.

For computational purposes, we recast MIND in terms of simpler unconstraint optimization problems via Lagrangian penalization as well as Fenchel duality. Performance of several MINDs is demonstrated on numerical examples.

Keywords: Fenchel duality, Lagrangian formulation, nonparametric regression, statistical imaging, change points, wavelets, high-dimensional linear models, variational estimation.

Contents

1 Introduction

Recovering a (typically high dimensional) parameter vector $\beta \in \mathbb{R}^p$ from (possibly indirect) noisy observations is a fundamental task in modern data analysis and has been a long standing topic of intense investigation in statistical, mathematical, and related sciences. Applications are vast. What we have primarily in mind are imaging and signal recovery problems, such as tomography or spectroscopy. Efficient recovery of these signals (we use the term *signal* in the following for a general β, including images as well) always relies on structural prior information, which is in many cases given by neighborhood information on the structure of the signal. In the context of imaging a neighborhood often is with respect to spatial distance, e.g. expressed in certain smoothness assumptions. In other scenarios (e.g. graphs) this should be understood as "structurally similar regions". Other assumptions may concern certain features such as peaks of a signal or texture of an image. Although a serial signal is of physical space dimension one and an image of dimension two or three, the mathematical "effective dimension" of the problem is given through the complexity of the modeling function system of the signal. For example, if a one-dimensional signal is sampled at n points and is assumed to be potentially different at each sampling point, the effective dimension p may be thought of $p = n$. Related to this are problems that are "truly" high dimensional in the sense that potentially many more coefficients (or parameters) have to be estimated than observations available, viz. $p \gg n$. Applications include high throughput data in various "omics" technologies in genetics or large scale networks, to mention a few.

Probably the most prominent and simplest unifying model in this context assumes a linear relationship $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ between the unknown coefficient vector $\beta \in \mathbb{R}^p$ and the observations $Y = (Y_1, \ldots, Y_n)$,

$$Y_i = (\boldsymbol{X}\beta)_i + \varepsilon_i \qquad \text{for } \ i = 1, \ldots, n\,. \tag{1}$$

Here n is the number of observations, p is the number of unknown parameters, Y_i are the observed data, and ε_i's model the error (noise) in the observations. Throughout this paper we assume for the sake of simplicity a white noise error, i.e. the ε_i's are independent and normally distributed random variables $\mathcal{N}(0, \sigma^2)$ with mean zero and variance σ^2. We stress, however, that much of the following can be extended to other models, e.g. to generalized linear exponential family regression, and to data with certain dependencies or heterogeneous errors.

Some convention of notation is as follows: Matrices and linear operators are in boldface. Vectors are in the column form and are written in normal font. For a vector $v = (v_i)_{i=1}^n$ in $\mathbb{R}^n$, we denote the ℓ^0-quasi-norm by $\|v\|_0 = \#\{i : v_i \neq 0, i = 1, \ldots, n\}$, the ℓ^2-norm by $\|v\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$, and the ℓ^∞-norm by $\|v\|_\infty = \max_{1\le i\le n} |v_i|$. In case of $n = p$, i.e. the number of observation equals the number of parameters, we always use n in place of p. Let $x_+ = \max\{x, 0\}$ denote the positive part of a real variable x.

1.1 The linear model: Examples

The linear model in (1) contains many important instances of statistical models. Concrete examples we shall consider are as follows.

(I) Non-parametric regression. Here we take $p = n$, $\boldsymbol{X} = \boldsymbol{I}$ as the n-dimensional identity matrix, and the unknown parameter

$$\beta = \boldsymbol{S}_n(f) := (f(x_i))_{i=1}^n$$

as values of a regression function $f\colon [0,1]^d \to \mathbb{R}$ at sampling points $x_i \in [0,1]^d$ for some dimension $d \ge 1$ (here the d-dimensional unit cube is chosen for simplicity). Given the vector β, the full regression function f can be recovered by some interpolation or approximation procedure that relies on additional structural assumptions on f, encoded in a function space $\mathcal{F}$, such as certain smoothness properties. These can be often expressed conveniently in terms of approximation properties with respect to a certain function system, such as radial basis functions (Wendland, 2005), splines (Wahba, 1990), polynomials, trigonometric series (Walter and Shen, 2001), or wavelets (Mallat, 2009). Many estimators do not explicitly reflect this decomposition and estimation of β and interpolation are performed simultaneously.

(II) Linear inverse problems. In this example, the unknown $\beta = \boldsymbol{S}_p\,(f) = (f(x_j))_{j=1}^p$ consists of values of an element in some function space $\mathcal{F}$ as in model (I). It is linked to the observational vector by a *system matrix*

$$\boldsymbol{X} = \boldsymbol{S}_n \circ \boldsymbol{K} \circ \boldsymbol{E}_p \colon \mathbb{R}^p \to \mathbb{R}^n,$$

which is the composition of an abstract interpolation operator $\boldsymbol{E}_p \colon \mathbb{R}^p \to \mathcal{F}$, a linear operator $\boldsymbol{K} \colon \mathcal{F} \to \mathcal{U}$ modeling the particular inverse problem, and another sampling operator $\boldsymbol{S}_n \colon \mathcal{U} \to \mathbb{R}^n$ (with a slight abuse of notation) on the model space $\mathcal{U} \supset \boldsymbol{K}(\mathcal{F})$. Prominent examples include the Radon transform in computed tomography (Donoho, 1995b, Natterer and Wübbeling, 2001), the Fourier transform in magnetic resonance tomography (Epstein, 2008) or convolution with a point spread function in optics (Bertero et al., 2009, Aspelmeier et al., 2015), ranging from astronomical imaging to high resolution microscopy. For $\boldsymbol{K}$ the identity, we obtain the regression model (I).

(III) Change point detection. We consider $f \colon [0,1] \to \mathbb{R}$ as in model (I) with $d = 1$, and take $\beta_i = f\,(x_i) - f\,(x_{i-1})$, for $i \in \{2, \ldots, n\}$, as the jump sizes of the piecewise constant regression function f sampled at the locations x_i. Further we denote by $\beta_1 = f\,(x_1)$ its offset. The function f then can be recovered from $\beta = (\beta_i)_{i=1}^n$ by $f\,(x_i) = \sum_{k=1}^i \beta_k$ for $i = 1, \ldots, n$. The relation between the jump sizes β and data Y_i can be written in the form of (1), where the system matrix $\boldsymbol{X}$ takes the form (Boysen et al., 2009)

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Taking $p = n$ reflects the fact that no a-priori assumptions on the number and locations of the jumps in this model have been posed. Neighboring information amounts to information on length of connected segments, i.e. how non-zero β_i's are located.

(IV) High dimensional regression. Here the unknown $\beta \in \mathbb{R}^p$ is a high dimensional parameter vector, typically $p \gg n$, which means that the number of unknown parameters is much larger than the number of observations. In contrast to the first two examples, in general, no neighboring structure on β is posed, rather a sparsity assumption (only a few of the p coefficients β_i are nonzero) that constrains the set of possible solutions (see e.g. Bühlmann and van de Geer, 2011, Wainwright, 2019 and the references therein).

Lastly, we stress that many models can be cast in the form of (1), but the error may not be an independent Gaussian white noise (thus beyond the scope of this paper). For example, consider the errors-in-variable (or measurement error) model given by

$$Y_i = (\boldsymbol{X}\beta)_i + \varepsilon_i, \quad i = 1, \ldots, n, \qquad \text{and} \qquad \tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E} \;\in\; \mathbb{R}^{n \times p}, \tag{2}$$

where we assume for simpilicity that ε_i's are independent $\mathcal{N}(0, \sigma^2)$ with variance σ^2, entries of $\boldsymbol{E}$ are independent $\mathcal{N}(0, \tau^2)$ with variance τ^2, and ε_i's and entries of $\boldsymbol{E}$ are independent. The aim is to recover the unknown parameter β given observations $Y = (Y_i)_{i=1}^n$ and $\tilde{\boldsymbol{X}}$ (which is a perturbed version of $\boldsymbol{X}$). (2) can be rewritten as

$$Y_i = (\tilde{\boldsymbol{X}}\beta)_i + \tilde{\varepsilon}_i \qquad \text{with} \quad \tilde{\varepsilon}_i = \varepsilon_i - (\boldsymbol{E}\beta)_i.$$

Note that $\tilde{\varepsilon}_i$'s are still independent and Gaussian, more precisely, $\mathcal{N}(0, \sigma^2 + \tau^2 \|\beta\|_2^2)$, but their variance now depend on the unknown parameter β (see e.g. Carroll et al., 2006).

1.2 Types of variational estimation

We introduce here two prominent types of variational methods for estimating the parameter β in (1), and examples will be given in subsequent sections.

Among the best known approaches is the method of maximum likelihood, which boils down to least squares estimation in our setup. It minimizes the residual sum of squares

$$C_Y(\beta) := \|\boldsymbol{X}\beta - Y\|_2^2 = \sum_{i=1}^{n} |(\boldsymbol{X}\beta)_i - Y_i|^2 \ . \tag{3}$$

If β is a low dimensional vector, the maximum likelihood estimator is well known to be asymptotically normal and efficient under proper regularity conditions on $\boldsymbol{X}$ (e.g. van der Vaart, 1998). Note, however, that model (III) is a notable exception (see Korostelev and Korosteleva, 2011).

In case that β represents many degrees of freedom, minimizing the residual sum of squares leads to over-fitting of the data. See e.g. Portnoy (1988) for conditions on p when the maximum likelihood estimator fails to be asymptotically normal. In a certain sense, this is even true, when β is assumed to have only a few nonzero coefficients (sparseness) because the model selection error is not uniformly controllable (Leeb and Pötscher, 2006), without further assumptions. As the system matrix $\boldsymbol{X}$ becomes more ill-conditioned (i.e. one or more eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$ are close to zero), all this becomes even more critical because the fluctuations of β are heavily damped through $\boldsymbol{X}$ and the reconstruction process becomes more unstable (see O'Sullivan, 1986 for some examples).

Hence, any reasonable estimation procedure for a high dimensional vector β has, either implicitly or explicitly, to account for additional properties of the unknown parameter, such as smoothness, sparsity, or other structural information (recall models (I)–(IV)), i.e. to *regularize* the solution. Such a-priori information can be incorporated by requiring the value of a *regularization functional* $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ to be small at the particular estimate. In the following, we discuss two types of approaches, at a first glance, seemingly unrelated. We are aware that our treatment is incomplete. For example, there is a rapidly growing literature on (nonparametric) Bayesian techniques to advise such regularization as well (see e.g. Ghosal and van der Vaart, 2017), which is, however, beyond the scope of this survey.

1.2.1 Penalized estimation

Probably the most prominent approach to include such structural information on β into the estimation process is to incorporate a *regularization functional* R into a *data fidelity term* $G(\boldsymbol{X}\beta; Y)$ that measures the discrepancy from the data, typically as an additive penalty. The penalized estimator results from a solution of the Lagrangian variational problem

$$\min_{\beta\in\mathbb{R}^p} G(\boldsymbol{X}\beta; Y) + \gamma R(\beta), \tag{4}$$

for some $\gamma > 0$. In case of $G(\boldsymbol{X}\beta; Y) = C_Y(\beta)$ in (3), this is known as the *penalized least squares* estimator, which is a particular case of the log-likelihood function of the model. General G results in e.g. *penalized maximum-likelihood* estimation (see e.g. Eggermont and LaRiccia, 2009). Solution of (4) yields regularized estimates with regularity measured in terms of the regularization functional R. A plethora of estimators follow this paradigm, some instances of R will be discussed later on.

The proper choice of penalty parameter γ in (4) relies on the precise structure of the signal, which is not available in practice. Thus, a data driven strategy for the selection of γ is needed, and turns out to be a challenging and delicate issue. A wide range of methods have been suggested in the literature. Details are far beyond the scope of this paper, so we just mention Mallows' C_p (Mallows, 2000, Li and Werner, 2020), cross-validation (Allen, 1974, Stone, 1974), generalized cross-validation (Wahba, 1977), plug-in techniques (Loader, 1999), bootstrap based methods (Breiman, 1992, Shao, 1996), and techniques that are built on the Lepskiĭ balancing principle (Lepskiĭ, 1990, Lepski et al., 1997), to name only a few. However, in the following, we present a constrained formulation of (4), which offers a statistically simpler strategy of the corresponding threshold parameter selection and thereby circumvents the difficulties encountered with the choice of γ to some extent.

1.2.2 Constrained estimation

A seemingly different approach to incorporate R into the estimation process is based on the idea to use the data fidelity term as a constraint, resulting in the *constrained* estimator

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^p} & R(\beta) \\ \text{s. t.} & G(\boldsymbol{X}\beta;Y) \le q\ , \end{cases} \tag{5}$$

for some threshold $q > 0$. It reduces to the *constrained least squares* estimator when $G(\boldsymbol{X}\beta;Y) = C_Y(\beta)$ in (3). In fact, it is easily seen and well known that constrained estimation is closely related to penalized estimation in (4), see Section 4. However, even in the simple case of $G(\boldsymbol{X}\beta;Y) = C_Y(\beta)$, the correspondence between the two parameters γ in (4) and q in (5) is not given explicitly and depends on the data Y. It is exactly the lack of this explicit correspondence that makes the different nature of these estimators, as the (data driven) choice of γ in (4) and q in (5) is difficult to "translate" from the constrained formulation into the penalized one and vice versa. As mentioned, the choice of q in (5) is comparably easier than that of γ in (4), because a proper choice of q depends essentially on the distribution property of the error ε_i in (1), which is usually (approximately) available. One exemplary method is Morozov's discrepancy principle (Morozov, 1966), see also Section 2.

In the following, some prototypical estimators will be introduced.

1.3 Examples and spatial adaptation

In this section, we work with the general model in (1), unless explicitly specified.

1.3.1 Smoothing splines

We start with a prominent instance of the penalized least squares estimator, a *smoothing spline* (Wahba, 1990). The regularization functional R is the squared ℓ^2-norm of the (discrete) derivative,

$$R(\beta) = \frac{1}{2}\sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|^2\ ,$$

or some higher order analog. Increasing the threshold q in (5) (or equivalently the penalty γ in (4)) yields smoother estimates (i.e. solutions of (4) and (5)) and therefore its particular choice sensitively affects the regularity of the resulting estimator.

1.3.2 Spatial adaptation

Recall that $\beta = \boldsymbol{S}_p(f) = (f(x_i))_{i=1}^p$. The parameters γ and q in the spline approach in Section 1.3.1 act globally over the domain of f but the regularity of a regression function f is often "spatially" varying, i.e. it will depend on $x \in [0,1]^d$. This suggests that better results can be obtained by introducing *local weights* $w_i > 0$ depending on the spatially varying smoothness, i.e.

$$R_w(\beta) = \frac{1}{2}\sum_{i=1}^{p-1} w_i \left|\beta_{i+1} - \beta_i\right|^2 .$$

Because the local smoothness of the underlying function is unknown one faces the problem of how actually choosing the weights w_i in an adaptive (data driven) manner. This turns out to be a difficult task as the minimization problems in (4) and (5) when additionally optimizing over w are not convex anymore, in general. Further, identification of the actual estimator for β and the weights w_i is in general not possible. A similar comment applies to the attempt to "localize" other global regularization functionals such as the total variation, although some methods have been suggested (e.g. Davies and Kovac, 2001, Dong et al., 2011, Lenzen and Berger, 2015). In the following we will therefore give a reformulation of this attempt in a more general context that avoids these difficulties and provides feasible estimators.

1.3.3 Wavelet soft-thresholding

Wavelet (and more generally dictionary) based thresholding methods have been proven to provide certain spatial adaptivity without explicitly including spatially varying weights (e.g. Donoho and Johnstone, 1994, 1995). Heuristically speaking, the reason is that the spatial variability has already been incorporated in the basis (or in general dictionary) functions and is respected by the thresholding procedure. This is an important feature, not shared by other series estimators, such as Fourier estimators, which are not localized in time domain (see e.g. Hart, 1997, Walter and Shen, 2001, Tsybakov, 2009). As an introductory example it is illustrative to represent the soft-thresholding wavelet estimator as a constraint estimator as in Donoho (1995a). Assume the nonparametric regression model (I), i.e. $\boldsymbol{X} = \boldsymbol{I}$. In its constraint formulation the coefficient vector of the soft-thresholded wavelet estimator is obtained as the (unique) minimizer of $\|\beta\|_2$ among all $\beta \in \mathbb{R}^n$ satisfying the constraint

$$\max_{\lambda \in \Lambda} \left|\langle \phi_\lambda, Y - \beta \rangle\right| \le q \qquad (6)$$

(see Theorem 3 in Section 3). Here $(\phi_\lambda)_{\lambda\in\Lambda}$ denotes a system of wavelets or some other system spanning $\mathbb{R}^n$ (see Section 3 for examples). This optimization problem obviously falls in the framework of (5). The residuals $Y - \beta$ are analyzed by the ℓ^∞-norm (i.e. the maximum of absolute values in (6)) in the wavelet domain.

1.3.4 The Dantzig selector

The *Dantzig selector* (Candès and Tao, 2007) for (1) is defined as a solution of the optimization problem

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^p} & \sum\limits_{j=1}^{p} |\beta_j| \\ \text{s.t.} & \max\limits_{1\le i\le p} |\langle \boldsymbol{X}_i, Y - \boldsymbol{X}\beta\rangle| \le q\ , \end{cases} \tag{7}$$

with $\boldsymbol{X}_i$ denoting the i-th column of $\boldsymbol{X}$. Under the assumption that $\boldsymbol{X}$ satisfies the *restricted isometry property* and β is *sparse*, Candès and Tao (2007) showed that $\|\hat\beta - \beta\|_2^2$ is, with high probability, bounded by a logarithmic quantity times the oracle risk $\sum_{i=1}^{p} \min\{\beta_i^2, \sigma^2\}$.

The Dantzig selector acts on scales which depend on the system matrix $\boldsymbol{X}$ itself. It is therefore not necessarily multiscale in nature: For example in case that $\boldsymbol{X}$ is the identity matrix (model (I)) only the smallest scales are taken into account, i.e. the constraint acts only on each single observation. Note the difference to the constrained least squares estimators where the side constraint acts only on the largest scale. Hence, these estimators measure the data fidelity in a complementary way from a statistical point of view. Both estimators can be extended to a truly multiscale estimator, which additionally takes all intermediate scales between these two extremes into account, as we will see in Section 2.

1.3.5 The lasso

The *least absolute shrinkage and selection operator* (lasso; Tibshirani, 1996) has been introduced as a constraint estimator, namely the solution of

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^p} \|Y - \boldsymbol{X}\beta\|_2^2 \\ \text{s.t.} \sum\limits_{i=1}^{p} |\beta_i| \le c\ . \end{cases}$$

Note that this is the converse formulation of the constraint estimator in (5). From this perspective, the Dantzig selector in (7) is a "reverse lasso" with data ℓ^2 fidelity term replaced by the ℓ^∞ norm (Bickel et al., 2009). We will investigate this in more detail later on.

1.3.6 Nemirovskii's estimator

Nemirovskii (1985) introduced for the nonparametric regression model a particular constrained estimator which, as the lasso, in a sense has reversed roles of the constraint and the objective in (5). This is done in the context of Sobolev spaces $W^{k,q}$, $k \in \mathbb{N}, 1 \le q \le \infty$. In the discrete setting of model (I) this reads as

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^n} \|Y - \beta\|_{\mathcal{N}} \\ \text{s.t. } \|\beta\|_{k,q} \le c\ . \end{cases} \tag{8}$$

Here $\|\beta\|_{k,q}$ is a discretized version of the d-dimensional (k,q) Sobolev norm $\|f\|_{W^{k,q}} := \sum_{0\le|l|\le k} ||D^l f||_{L^q}$ and $\beta = \boldsymbol{S}_n(f) = (f(x_i))_{i=1}^n$. The norm $\|\cdot\|_{\mathcal{N}}$ is a multiscale analog to the ℓ^∞-norm and defined as

$$\|\beta\|_{\mathcal{N}} := \sup_{B\in\mathcal{N}} \frac{1}{\sqrt{|B|}} \left|\sum_{i\in B} \beta_i\right|$$

for a system $\mathcal{N}$ of sets $B \subset \{1, \cdots, n\}$. As suggested by Nemirovskii (1985), $\mathcal{N}$ is called *normal* if it obeys a certain covering property, e.g. the system of all subsquares or discrete balls does satisfy such a condition. In fact, normal systems of cardinality $\mathcal{O}(n)$ exist (Grasmair et al., 2018).

1.4 Outline of this paper

The major aim of this paper is to unify these estimators (and extensions, which we will discuss later on) and shed some light on their commonalities and differences from a variational point of view. This allows for some better statistical understanding but also serves as a guide for a unifying algorithmic treatment. We stress that most of this is known and scattered over the literature in different contexts and communities. However, we are not aware of a comprehensive and unifying approach for all these estimators from the view point of optimization characterization.

The rest of the paper is organized as follows. In Section 2, the general class of "MIND" (MultIscale Nemirovskii–Dantzig) estimators is introduced, which comprises all of the aforementioned estimators and generalizations thereof. In the course of Sections 3 and 4 selective instances of MIND are discussed. This includes, for example, various (block) thresholding strategies, multiscale extensions of the Dantzig selector, the group lasso, and (reverse) Nemirovskii's estimator. The distributional properties of the multiscale constraint of MIND, are summarized in Section 5. The implications for the selection of a proper threshold for MIND are also discussed there. In Section 6 various algorithms for the computation of MIND are discussed, based on which several numerical examples are presented as illustrations. All proofs are deferred to the Appendix.

2 MIND: The multiscale Nemirovskii–Dantzig estimator

Classical variational methods, such as penalized or constrained least squares, control the residual vector $Y - \boldsymbol{X}\beta$ in a global way (thus on a single scale only), in general. This is in contrast to dictionary based multiscale methods, like wavelet estimators.

The MIND estimator introduced in Definition 1 below can be seen as a hybrid approach between variational methods and dictionary (e.g. wavelet) based multiscale methods. It follows the philosophy of dictionary based methods to analyze the residual vector simultaneously over a whole family of scales and locations. At the same time it does not necessarily rely on an explicit dictionary expansion for its regularization term. Instead it allows for a more general variational regularization formulation to employ smoothness information about the unknown β in terms of the regularization functional R.

Definition 1 (MultIscale Nemirovskii–Dantzig estimator, MIND; Grasmair et al., 2018)**.** *For a given index set $\mathcal{A}$, let $(\boldsymbol{\Pi}_a)_{a\in\mathcal{A}}$ be a family of linear transformations*

$$\boldsymbol{\Pi}_a \colon \mathbb{R}^n \to \mathbb{R}^{n_a}\,, \qquad \text{for } a \in \mathcal{A}\,,$$

$(w_a)_{a\in\mathcal{A}}$ a family of positive numbers and $(s_a)_{a\in\mathcal{A}}$ a family of nonnegative numbers. Let also

$R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ *be a functional. Any solution of the constrained optimization problem*

$$\begin{cases} \min_{\beta \in \mathbb{R}^p} & R(\beta) \\ \text{s. t.} & \max_{a \in \mathcal{A}} \left\{ \frac{\|\mathbf{\Pi}_a (Y - \boldsymbol{X}\beta)\|_2}{w_a} - s_a \right\} \leq q \end{cases} \tag{9}$$

is called a MIND *for* (1) *with regularizer* R*, probe functionals* $(\mathbf{\Pi}_a)_{a\in\mathcal{A}}$*, weights* $(w_a)_{a\in\mathcal{A}}$*,* $(s_a)_{a\in\mathcal{A}}$ *and threshold* $q > 0$*.*

Obviously, the Dantzig selector in (7) is a special instance of (9) as the columns of X can be absorbed into the matrix $\mathbf{\Pi}_a$; Also $w_a = 1$, $s_a = 0$ and $R(\beta)$ equals the (one-dimensional) discrete total variation. Many other examples will be given later.

The weights w_a could be absorbed into $\mathbf{\Pi}_a$, but, because they often play a particular role as scale factors, we do not. Further, the penalty s_a is included in order to balance the random perturbations caused by noise over different scales. This balancing idea was formalized by Dümbgen and Spokoiny (2001), see also Walther and Perry (2020) and the references therein. The proper choices of w_a and s_a depend on the type of signal preferably to be reconstructed and the model itself, see e.g. Schmidt-Hieber et al. (2013) for convolution models, and Spokoiny (2009), Frick et al. (2014), Pein et al. (2017) and Li et al. (2019) for change point regression. The constraint in (9) forces the residuals $Y - \boldsymbol{X}\beta$ to satisfy

$$\|\mathbf{\Pi}_a (Y - \boldsymbol{X}\beta)\|_2^2 = \sum_{i=1}^{n_a} |(\mathbf{\Pi}_a (Y - \boldsymbol{X}\beta))_i|^2 \leq (q + s_a)^2 w_a^2 \,,$$

simultaneously for every $a \in \mathcal{A}$. The threshold q determines the size of the feasibility region and acts as a tuning parameter: the larger q the more the constraint in (9) is relaxed and hence the smoother (measured in terms of the functional R) the MIND will be.

The collection of probe functionals $(\mathbf{\Pi}_a)_{a\in\mathcal{A}}$ encodes *multiple scales*: For instance, if $\mathbf{\Pi}_a v = \langle \phi_a, v\rangle$ for some vector $\phi_a \in \mathbb{R}^n$, namely, $n_a = 1$, then the size of the support of ϕ_a (i.e. the number of nonzero entries of ϕ_a) is interpreted as its "scale". Probe functionals are often motivated by examining whether there is remaining structure left in the residual $Y - \boldsymbol{X}\beta$ for a candidate β. This relates to the detection of anomalies (hot spots) in (spatial) random fields, see e.g. Sharpnack and Arias-Castro (2016) and Proksch et al. (2018) for Gaussian errors, and König et al. (2020) for extension to non-Gaussian models. Roughly, the detection of signal is formalized as a multiple testing problem for the collection of hypotheses

$$H_a \, \mathbf{\Pi}_a (Y - \boldsymbol{X}\beta) \text{ contains purely noise,} \qquad \text{for } \; a \in \mathcal{A}\,.$$

In this sense, the constraint in (9) is interpreted as the acceptance region of a multiple test, and thus every MIND aims at finding the most regular candidate measured by R within this acceptance region. That is, MIND can be seen as *a combination of multiple testing and variational estimation.* This does not only provide guidance in designing the probe functionals $\mathbf{\Pi}_a$, but also suggest rules for the choice of threshold q. For instance, a reasonable rule is to control the *familywise error rate*, which is the probability of making any wrong rejections. It suggests to select q such that

$$\inf_{\beta\in\mathbb{R}^p} \mathbf{P}_\beta \left\{ \|\mathbf{\Pi}_a (Y - \boldsymbol{X}\beta)\|_2^2 \leq (q + s_a)^2 w_a^2 \quad \text{for all } \; a \in \mathcal{A} \right\} \geq 1 - \alpha \tag{10}$$

for some error level $\alpha \in (0,1)$. As the value of q is independent of β, Definition 1 readily implies the *simultaneous statistical guarantee*

$$\inf_{\beta \in \mathbb{R}^p} \mathbf{P}_\beta \left\{ R(\hat{\beta}) \leq R(\beta) \right\} \geq 1 - \alpha, \tag{11}$$

where $\hat{\beta}$ denotes the MIND. Note that the objective R in (9) ensures the MIND to fulfill certain desired regularity properties. Such a statistical guarantee reveals a statistical balancing between data approximation and regularity of β. The data fidelity constraint of MIND in (9) enforces the closeness to the data, but at the same hand the minimization of R provides also smoothness control: (11) says that, independent of the true parameter β, the MIND is no less regular (in terms of R) than the true parameter with probability at least $1 - \alpha$.

The quantile $q \equiv q_\alpha$ can be estimated via Monte Carlo simulations (see e.g. Frick et al., 2012). Besides it can be approximated using the limiting distribution of the *multiscale statistic* (recall (9))

$$T_n \;\equiv\; \max_{a \in \mathcal{A}} \left\{ \frac{\|\mathbf{\Pi}_a \left(Y - \boldsymbol{X}\beta\right)\|_2}{w_a} - s_a \right\} \;=\; \max_{a \in \mathcal{A}} \left\{ \frac{\|\mathbf{\Pi}_a \, \varepsilon\|_2}{w_a} - s_a \right\}, \tag{12}$$

which will be discussed in Section 5. Such choices of q turn out to be extremely favorable in practice, since we do not require any knowledge of the true parameter, or to estimate it from the data, in contrast to the penalized or the constrained optimization formulation with switching roles of objective and constraint.

In the sequel we consider several further estimation and thresholding techniques that are shown to be particular examples of the MIND.

3 Thresholding methods

In this section we relate the MIND principle to so-called thresholding based estimators for (1) which rely on an (explicit) expansion of β. Most thresholding techniques have been initially designed for solving the nonparametric regression problem (I). An extension to the linear inverse problem is the wavelet–vaguelette decomposition, which will be discussed in Section 3.5 in more detail.

3.1 Wavelet soft-thresholding

Let $(\phi_\lambda)_{\lambda \in \Lambda}$ be an orthonormal (discrete) wavelet basis of $\mathbb{R}^n$, i.e. $\|\phi_\lambda\|_2 = 1$ and $\langle \phi_\lambda, \phi_{\lambda'} \rangle = 0$ for $\lambda \neq \lambda'$, see for example Vidakovic (1999) and Mallat (2009). Due to orthonormality, every parameter vector $\beta \in \mathbb{R}^n$ can be uniquely expanded in the wavelet basis

$$\beta = \sum_{\lambda \in \Lambda} \langle \phi_\lambda, \beta \rangle \, \phi_\lambda \, .$$

Example 2 (Haar wavelets)**.** *The Haar father wavelet (or scaling function)* $\phi^{(\mathrm{H})}$ *is defined as*

$$\phi^{(\mathrm{H})}(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and the Haar mother wavelet (or wavelet function) as $\psi^{(\mathrm{H})}(x) := \phi^{(\mathrm{H})}(2x) - \phi^{(\mathrm{H})}(2x-1)$ *for* $x \in \mathbb{R}$*. Let* $\psi^{(\mathrm{H})}_{j,k}(x) = 2^{j/2}\psi^{(\mathrm{H})}(2^j x - k)$ *for* $x \in \mathbb{R}$*, where* $j \in \mathbb{N}_0$ *corresponds to the* scale *and*

$k \in \mathbb{N}_0$ to the location *of $\psi^{(\mathrm{H})}_{j,k}$. For simplicity, we consider $n = 2^J$ with some $J \in \mathbb{N}_0$. It is known that*

$$\mathcal{B} \equiv \left\{ \phi^{(\mathrm{H},n)},\, \psi^{(\mathrm{H},n)}_{j,k} \ :\ j = 0, \ldots, J-1 \text{ and } k = 0, \ldots, 2^j - 1 \right\}$$

forms an orthonormal basis of $\mathbb{R}^n$ (see e.g. Mallat, 2009), where the included basis elements are defined by sampling the scaling function and the scaled wavelet functions,

$$\phi^{(\mathrm{H},n)} = \boldsymbol{S}_n\big(\phi^{(\mathrm{H})}\big) := \big(\phi^{(\mathrm{H})}(i/n)\big)_{i=0}^{n-1} \quad \text{and} \quad \phi^{(\mathrm{H},n)}_{j,k} = \boldsymbol{S}_n\big(\psi^{(\mathrm{H})}_{j,k}\big) := \big(\psi^{(\mathrm{H})}_{j,k}(i/n)\big)_{i=0}^{n-1}.$$

In our notation, the basis $\mathcal{B}$ will be re-indexed as $\{\phi_\lambda : \lambda \in \Lambda\}$.

Some other examples of orthonormal wavelets include Daubechies' wavelets, Coiflets, Meyer wavelets and B-spline wavelets, see e.g. Chui (1992) and Daubechies (1992).

For any $q > 0$, the nonlinear *soft-thresholding function* is given as $\eta^{(\mathrm{soft})}(\,\cdot\,, q) : \mathbb{R} \to \mathbb{R}$,

$$\eta^{(\mathrm{soft})}(x, q) := x\left(1 - \frac{q}{|x|}\right)_+ \quad \text{for } x \in \mathbb{R}\,.$$

The soft-thresholding function sets any coefficient x with $|x|$ below the threshold q to zero and shrinks the remaining coefficients towards zero by the constant value q. Moreover, one notices that $\eta^{(\mathrm{soft})}(\,\cdot\,, q)$ is a continuous function on the whole real line, see **Figure 1**.

This leads to one of the most prominent wavelet shrinkage methods, the *wavelet soft-thresholding estimator* (Donoho, 1995a)

$$\hat{\beta}^{(\mathrm{soft})} = \sum_{\lambda \in \Lambda} \eta^{(\mathrm{soft})}\left(\langle \phi_\lambda, Y \rangle, q_\lambda\right) \phi_\lambda\,. \tag{13}$$

Here the thresholds $q_\lambda > 0$, for $\lambda \in \Lambda$, are tuning parameters that determine whether or not an empirical wavelet coefficient $\langle \phi_\lambda, Y \rangle$ is accepted as a wavelet coefficient for the estimate $\hat{\beta}^{(\mathrm{soft})}$. In its elementary form, wavelet soft-thresholding is often used with the *universal threshold* $q_\lambda = q := \sigma\sqrt{2 \log n}$, which is independent of the index $\lambda \in \Lambda$, see Donoho (1995a) and Section 5. Extensions to scale or data dependent thresholds have been investigated in Donoho and Johnstone (1995) and Antoniadis and Fan (2001) among many others.

Using the fast wavelet transform algorithm (see e.g Daubechies, 1992, Cohen, 2003, Mallat, 2009) the wavelet thresholding estimator can be computed with only $\mathcal{O}(n)$ floating point operations. Despite the simplicity and computational efficiency, soft-thresholding has various statistical optimal recovery properties, such as adaptive minimax optimality over various Besov balls (Donoho and Johnstone, 1994, 1995).

The next result reveals soft-thresholding as an instance of the MIND in (9) with probe functionals $\mathbf{\Pi}_a = \mathbf{\Pi}_\lambda$ defined to map $v \in \mathbb{R}^n$ to $\langle \phi_\lambda, v \rangle$, and weights taken as $w_a = w_\lambda = q_\lambda$ and $s_a = 0$.

Theorem 3 (Variational characterization of soft-thresholding; Donoho, 1995a)**.** *Let $(\phi_\lambda)_{\lambda \in \Lambda}$ be an orthonormal basis of $\mathbb{R}^n$. For any $r \in (0, \infty)$, the soft-thresholding estimator $\hat{\beta}^{(\mathrm{soft})}$ defined by* (13) *is the unique solution of*

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^n} & \sum\limits_{\lambda \in \Lambda} |\langle \phi_\lambda, \beta \rangle|^r \\ \text{s. t.} & \max\limits_{\lambda \in \Lambda} \dfrac{|\langle \phi_\lambda, Y - \beta \rangle|}{q_\lambda} \leq 1\,. \end{cases}$$

The same holds true, if we replace the objective function above by $\|\beta\|_2^2 = \sum_{i=1}^n \beta_i^2$.

Note that the variational characterization of soft-thresholding in Theorem 3 holds for any orthonormal basis $(\phi_\lambda)_{\lambda\in\Lambda}$, e.g. a Fourier basis in place of the wavelet basis. However, soft-thresholding is often applied with a wavelet basis (or a similar multiscale system) since its optimality properties for function estimation heavily depends on the multiscale structure and the spatial adaptivity of wavelets (Donoho, 1995a).

Wavelet soft-thresholding can further be equivalently characterized as the unique solution of the penalized least squares functional (cf. (4))

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|Y-\beta\|_2^2 + \sum_{\lambda\in\Lambda} q_\lambda \left|\langle \phi_\lambda, \beta\rangle\right| ,$$

where the penalty R is taken as the ℓ^1-norm of the wavelet coefficients of the parameter vector (see Donoho, 1995a, Chambolle et al., 1998). This characterization of soft-thresholding again can be verified in an elementary manner following the proof of Theorem 3. Here the equivalence of the penalized and the constrained problem is, however, not accidental. In fact, this equivalence is a special case of Theorem 8 below and due to Fenchel duality (see Appendix A.2).

3.2 Modified thresholding nonlinearity

The soft-thresholding function $\eta^{(\mathrm{soft})}(\,\cdot\,, q)$ systematically shrinks coefficients towards zero, at least if the thresholds are taken to be independent of the data. This yields a finite sample bias if there is a significant amount of non-zero coefficients. To overcome this issue, various modified thresholding functions have been proposed in the literature.

- Hard-thresholding function (Donoho and Johnstone, 1994).
  Besides the soft-thresholding function, the *hard-thresholding function*,

  $$\eta^{(\mathrm{hard})}(x,q) := \begin{cases} 0 & \text{if } |x| \le q \\ x & \text{otherwise} \end{cases} ,$$

  is the most basic and best known thresholding function. The hard-thresholding function does not shrink large coefficients and therefore usually yields a smaller mean square error than soft-thresholding. However, the hard-thresholding function has discontinuities at $x = \pm q$, which sometimes yields visually disturbing artifacts for signal and image recovery.

- Nonnegative garrote (Breiman, 1995, Gao, 1998).
  The *nonnegative garrote thresholding function* is a one-dimensional version of the well known James–Stein shrinkage function (James and Stein, 1961), and is defined by

  $$\eta^{(\mathrm{JS})}(x,q) := x\left(1-\frac{q^2}{|x|^2}\right)_+ .$$

  It is continuous and has a vanishing shrinkage effect as $|x|$ tends to infinity. Therefore, it is often claimed to combine the advantages of soft- and hard-thresholding.

All thresholding functions defined above are special cases of the class of functions

$$\eta_\theta(x,q) := x\left(1-\frac{q^\theta}{|x|^\theta}\right)_+$$

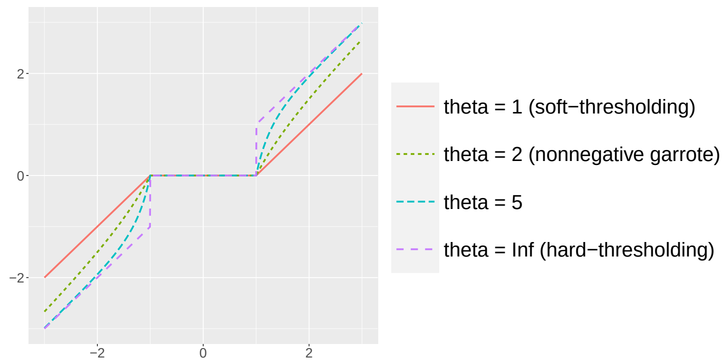

Figure 1: The thresholding functions $\eta_\theta(\,\cdot\,, q)$, plotted for the threshold $q = 1$, are antisymmetric and equal to zero on $\{x : |x| \leq q\}$. It coincides with the soft-thresholding for $\theta = 1$, with the nonnegative garrote for $\theta = 2$, and converges pointwise and monotonically to the hard-thresholding for $\theta \to \infty$.

for some specific value of $\theta > 0$. In fact, the soft-thresholding function corresponds to $\theta = 1$, the nonnegative garrote to $\theta = 2$, and the hard-thresholding function to the limiting case $\theta \to \infty$. See again **Figure 1**.

We observe that the soft-thresholding function and the nonnegative garrote shrinkage function are related via the explicit relation

$$\eta^{(\mathrm{JS})}(x, q) = \eta^{(\mathrm{soft})}\left(x, \frac{q^2}{|x|}\right) \quad \text{for } x \neq 0\,.$$

(A similar relation, of course, holds for any of the thresholding functions η_θ.) This basic identity allows to interpret shrinkage by the nonnegative garrote as soft-thresholding applied with the threshold $q^2/|x|$, which is now data dependent. Based on this simple observation, one can carry over many properties of the soft-thresholding estimator to the *nonnegative garrote (or James–Stein) estimator*

$$\hat{\beta}^{(\mathrm{JS})} = \sum_{\lambda \in \Lambda} \eta^{(\mathrm{JS})}\left(\langle \phi_\lambda, Y \rangle, q_\lambda\right) \phi_\lambda\,.$$

Theorem 4 (Variational characterization of the nonnegative garrote)**.** *Assume the setting of Theorem 3. For any $r \in (0, \infty)$, the nonnegative garrote estimator $\hat{\beta}^{(\mathrm{JS})}$ defined above is the unique solution of*

$$\begin{cases} \min_{\beta \in \mathbb{R}^p} & \sum_\lambda |\langle \phi_\lambda, \beta \rangle|^r \\ \text{s.t.} & \max_{\lambda \in \Lambda} \dfrac{|\langle \phi_\lambda, Y - \beta \rangle|}{w_\lambda} \leq 1\,. \end{cases}$$

Here $w_\lambda = q_\lambda^2 / \max\{q_\lambda, |\langle \phi_\lambda, Y \rangle|\}$ are weights depending on the data coefficients $\langle \phi_\lambda, Y \rangle$ and the thresholds q_λ. The same holds true, if we replace the objective function by $\|\beta\|_2^2$.

The optimization problem in Theorem 4 is obviously an instance of the MIND in (9), with objective $R(\beta) = \sum_\lambda |\langle \phi_\lambda, \beta \rangle|^r$, probe functionals $(\langle \phi_\lambda, \cdot\,\rangle : \lambda \in \Lambda)$, and data dependent weights $(w_\lambda \colon \lambda \in \Lambda)$. Allowing data dependent weights reveal almost any thresholding technique as a

MIND. Even the hard-thresholding estimator may be written as a solution of the minimization problem in Theorem 4 if one replaces the weights by

$$w_\lambda \equiv w_\lambda\left(\langle \phi_\lambda, Y\rangle, q_\lambda\right) = \begin{cases} q_\lambda & \text{if } |\langle \phi_\lambda, Y\rangle| \le q_\lambda \\ 0 & \text{if } |\langle \phi_\lambda, Y\rangle| > q_\lambda \,. \end{cases}$$

Here we set $0/0 = 1$ and $x/0 = \infty$ for $x > 0$. In some sense, hard-thresholding is a degenerate situation, where the reciprocal weights $1/w_\lambda$ become singular for $|\langle \phi_\lambda, Y\rangle| > q_\lambda$.

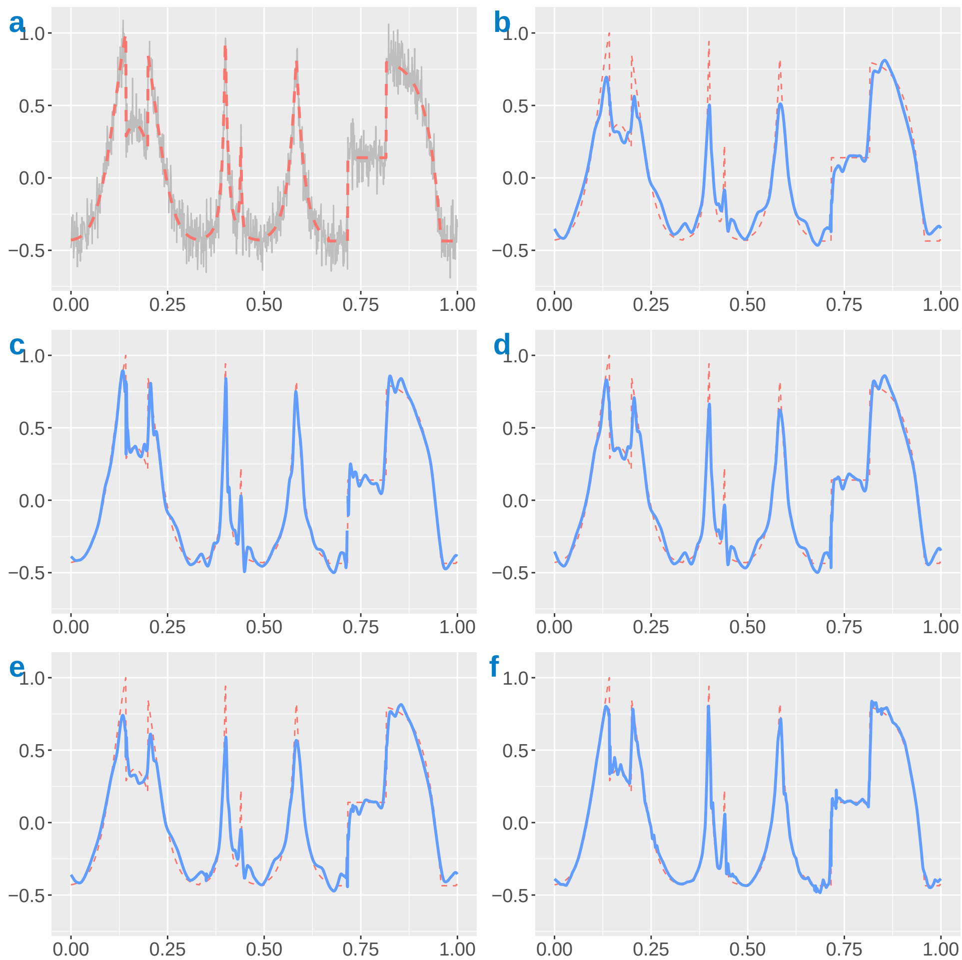

Figure 2: Regression of "piecewise smooth" signal via various thresholding strategies. (a) Noisy data (solid gray line) of 1,024 samples in model (I) with $\mathcal{N}(0, 0.1^2)$ error; (b)–(f) Estimators (solid blue line) of soft-thresholding, hard-thresholding, nonnegative garrote, FDR soft-thresholding (Abramovich and Benjamini, 1996) and block soft-thresholding, respectively. The true signal (dashed red line) is plotted in all panels. Daubechies' least asymmetric (symlets) with six vanishing moments are used.

We illustrate the performance of various thresholding methods in **Figure 2**. It shows that soft-thresholding shrinks sharp peaks while hard-thresholding introduces oscillating artifacts; Nonnegative garrote is a compromise between the two. Besides the choice of the nonlinear shrinkage function, the performance can be improved e.g. by an FDR (false discovery rate) strategy (Abramovich and Benjamini, 1996), or by a block-wise strategy (discussed below).

3.3 Block thresholding

In wavelet block thresholding the thresholding operation is not applied separately to each individual wavelet coefficient, but uniformly for a whole group of wavelet coefficients (e.g. Hall et al., 1997, Cai, 1999, Cai and Zhou, 2009, Chesneau et al., 2010). For that purpose, the set of all wavelet indices

$$\Lambda = \bigcup_{a \in \mathcal{A}} \Lambda_a$$

is grouped into several *disjoint* subsets $\Lambda_a \subset \Lambda$. One defines, for any $a \in \mathcal{A}$, blocks of wavelet coefficients

$$\mathbf{\Phi}_a Y := \left(\langle \phi_\lambda, Y \rangle : \lambda \in \Lambda_a \right) . \tag{14}$$

The wavelet thresholding is then performed uniformly for all coefficients within the same block. As in the case of component-wise thresholding, every particular block thresholding estimator depends on the thresholding function. A class of block thresholding function is

$$\eta_\theta(x, q) := x \left(1 - \frac{q^\theta}{\|x\|_2^\theta} \right)_+ \qquad \text{for} \qquad \theta > 0,$$

which now is applied to the whole block of coefficients, $x = (x_\lambda : \lambda \in \Lambda_a) \in \mathbb{R}^{\Lambda_a}$, instead of a single coefficient. The *block soft-thresholding estimator* and the *block James–Stein estimator*, respectively, are then defined by

$$\hat{\beta}^{(\text{b-soft})} = \sum_{a \in \mathcal{A}} \sum_{\lambda \in \Lambda_a} \left(\eta_1 \left(\mathbf{\Phi}_a Y, q_a \right) \right)_\lambda \phi_\lambda \,, \tag{15}$$

$$\hat{\beta}^{(\text{b-JS})} = \sum_{a \in \mathcal{A}} \sum_{\lambda \in \Lambda_a} \left(\eta_2 \left(\mathbf{\Phi}_a Y, q_a \right) \right)_\lambda \phi_\lambda \,,$$

where $q_a > 0$ are possibly block dependent thresholds. Similarly to Theorem 3, the block soft-thresholding estimator and the block James–Stein estimator are again special instances of the MIND in (9).

Theorem 5 (Variational characterization of block soft-thresholding)**.** *For any $r \in (0, \infty)$, the block soft-thresholding estimator $\hat{\beta}^{(b\text{-}soft)}$ in* (15) *is the unique solution of*

$$\begin{cases} \min_{\beta \in \mathbb{R}^p} & \sum_\lambda |\langle \phi_\lambda, \beta \rangle|^r \\ \text{s.t.} & \max_{a \in \mathcal{A}} \frac{\|\mathbf{\Phi}_a (Y - \beta)\|_\infty}{q_a} \leq 1 \,. \end{cases}$$

The same holds true if we replace the objective function above by $\|\beta\|_2^2 = \sum_i \beta_i^2$.

As in the non-block case, the block James–Stein thresholding $\eta^{(\text{b}-\text{JS})}(x, q)$ can be expressed in terms of block soft-thresholding applied with the threshold $q^2 / \|x\|_2$. Hence the block James–Stein estimator can be characterized by the minimization in Theorem 5 with q_a replaced by

$$w_a = w_a(\mathbf{\Phi}_a Y, q_a) = \frac{q_a^2}{\max\{q_a, \|\mathbf{\Phi}_a Y\|_2\}} \,.$$

The same again holds true if we replace the objective by the squared ℓ^2 norm $\|\beta\|_2^2$.

Recall that a wavelet basis is indexed by scales and locations, see Example 2. In wavelet block thresholding, each set Λ_a is usually supposed to have a fixed scale and to consist of an interval (or block) of location indices. We emphasize that for the above characterizations of the block thresholding methods such an additional restriction is not required. Further, the wavelet basis may be replaced by an arbitrary orthonormal basis.

3.4 Thresholding in frames

The rationale behind wavelet thresholding is that the parameter to be recovered can be efficiently represented as a sparse linear combination of elements of the wavelet basis. In real world signal and image processing applications the parameter β is usually not strictly sparse and the removal of small coefficients often introduces visually disturbing artifact (see **Figure 2** and e.g. Donoho and Johnstone, 1994, Coifman and Donoho, 1995, Donoho, 1995a, Candès and Donoho, 1999, Mallat, 2009, Starck et al., 2010, Grasmair et al., 2018).

One way of addressing this issue is to consider an overcomplete frame or dictionary instead of a wavelet basis. A *dictionary* of $\mathbb{R}^n$ is a family $(\phi_\lambda)_{\lambda\in\Lambda}$ of elements that spans the whole space $\mathbb{R}^n$. Hence any $\beta \in \mathbb{R}^n$ can be written in the form $\beta = \sum_{\lambda\in\Lambda} x_\lambda \varphi_\lambda$ for certain coefficients $x_\lambda \in \mathbb{R}$. If there are some constants $0 < a \leq b < \infty$ such that

$$a \|\beta\|_2^2 \leq \sum_{\lambda\in\Lambda} |\langle \phi_\lambda, \beta\rangle|^2 \leq b \|\beta\|_2^2 \quad \text{for all } \beta \in \mathbb{R}^n \,, \tag{16}$$

then $(\phi_\lambda)_{\lambda\in\Lambda}$ is called a *frame.* In such a situation, the mapping

$$\mathbf{\Phi} \colon \mathbb{R}^n \to \mathbb{R}^\Lambda \colon \beta \mapsto (\langle \phi_\lambda, \beta\rangle)_{\lambda\in\Lambda}$$

is called the *analysis operator* and $\mathbf{\Phi}^\mathsf{T}\,\mathbf{\Phi}$ the *frame operator.* Due to the frame property, the frame operator is invertible and the elements $\psi_\lambda = (\mathbf{\Phi}^\mathsf{T}\,\mathbf{\Phi})^{-1}\phi_\lambda$ are well defined. They again form a frame $(\psi_\lambda)_{\lambda\in\Lambda}$, which is called the *dual frame.* Note that in a finite dimensional situation any dictionary is automatically a frame, but this is not the case for highly redundant frames in infinite dimensional spaces.

If $(\phi_\lambda)_{\lambda\in\Lambda}$ is a frame, then one has the reproducing formula $\beta = \sum_{\lambda\in\Lambda} \langle \phi_\lambda, \beta\rangle\, \psi_\lambda$ where $(\psi_\lambda)_{\lambda\in\Lambda}$ its frame dual to $(\phi_\lambda)_{\lambda\in\Lambda}$. This motivates the following definition of a *frame based soft-thresholding estimator*

$$\hat{\beta}^{(\mathrm{soft})} = \sum_{\lambda\in\Lambda} \eta^{(\mathrm{soft})} \left(\langle \phi_\lambda, Y\rangle, q_\lambda\right) \psi_\lambda \,.$$

It can be written as $\hat{\beta}^{(\mathrm{soft})} = \sum_{\lambda\in\Lambda} \hat{x}_\lambda \psi_\lambda$ with $(\hat{x}_\lambda)_{\lambda\in\Lambda}$ being the unique minimizer of

$$\begin{cases} \min\limits_{x\in\mathbb{R}^\Lambda} & \sum\limits_{\lambda} |x_\lambda|^r \\ \text{s.t.} & \max\limits_{\lambda\in\Lambda} \dfrac{|\langle \phi_\lambda, Y\rangle - x_\lambda|}{q_\lambda} \leq 1 \end{cases} \qquad \text{for } 0 < r < \infty.$$

Hence the coefficients $\hat{x}_\lambda$ can be viewed as a special case of the MIND in (9).

Popular redundant system used for thresholding in signal and image processing are translation invariant wavelet systems (Coifman and Donoho, 1995, Nason and Silverman, 1995, Lang et al., 1996, Pesquet et al., 1996), curvelets (Candès and Donoho, 2000, Starck et al., 2002, Ma and Plonka, 2010), shearlets (Labate et al., 2005, Kutyniok et al., 2012), contourlets (Do and Vetterli, 2005), or needlets (Kerkyacharian et al., 2010). Its particular choice will depend on prior information of the signal, the underlying geometry of the domain or computational aspects. In general, all thresholding techniques from before (e.g. hard- or block thresholding) can be applied here, but as the coefficients now become dependent, there are some subtle differences to orthogonal systems, see also Section 5.

3.5 Wavelet–vaguelette decomposition and related approaches

The wavelet–vaguelette decomposition is introduced in Donoho (1995b) to generalize wavelet techniques from nonparametric regression to linear inverse problems. Recall model (II). Let $\boldsymbol{K}\colon \mathcal{F} \to \mathcal{U}$ be a bounded linear operator mapping the function space $\mathcal{F} \subseteq L^2(\Omega)$ to a Hilbert space $\mathcal{U}$.

Definition 6 (Wavelet–vaguelette decomposition, Donoho, 1995b)**.** *The family* $(\phi_\lambda, u_\lambda, v_\lambda, \kappa_\lambda)_{\lambda\in\Lambda}$ *is called a* wavelet–vaguelette decomposition *for the linear operator* $\boldsymbol{K}\colon \mathcal{F} \to \mathcal{U}$*, if the following* quasi-singular value decompositions *hold:*

$$\begin{aligned} \boldsymbol{K}\,\phi_\lambda &= \kappa_\lambda u_\lambda\,, \qquad \textit{for } \lambda \in \Lambda\,, \\ \boldsymbol{K}^*\,v_\lambda &= \kappa_\lambda \phi_\lambda\,, \qquad \textit{for } \lambda \in \Lambda\,, \end{aligned}$$

where $(\phi_\lambda)_{\lambda\in\Lambda}$ *is an orthonormal wavelet basis of* $\mathcal{F}$*,* $(u_\lambda)_{\lambda\in\Lambda}$ *and* $(v_\lambda)_{\lambda\in\Lambda}$ *are two bases of the space* $\mathcal{U}$ *such that* $\langle v_\lambda, u_{\lambda'}\rangle = \delta_{\lambda,\lambda'}$ *for all* $\lambda, \lambda' \in \Lambda$*, and* $(\kappa_\lambda)_{\lambda\in\Lambda}$ *is a family of nonnegative numbers, referred to as* quasi-singular values*, which only depend the scale index but not the spatial index of the wavelets* ϕ_λ*.*

If $(\phi_\lambda, u_\lambda, v_\lambda, \kappa_\lambda)_{\lambda\in\Lambda}$ is a wavelet–vaguelette decomposition for the operator $\boldsymbol{K}$, then we have the reproducing formula

$$f = \sum_{\lambda\in\Lambda} \frac{\langle v_\lambda, \boldsymbol{K}\,f\rangle}{\kappa_\lambda}\,\phi_\lambda \qquad \text{for } f \in \mathcal{F}\,.$$

In the case of noisy observations $g = \boldsymbol{K}\,f + \varepsilon$, with ε denoting a white noise process, the *wavelet–vaguelette soft-thresholding estimator* (Donoho, 1995b) for f is defined by

$$\hat{f}^{(\mathrm{WV})} = \sum_{\lambda\in\Lambda} \eta^{(\mathrm{soft})}\left(\frac{\langle v_\lambda, g\rangle}{\kappa_\lambda}, q_\lambda\right)\,\phi_\lambda\,,$$

with $q_\lambda > 0$ denoting certain scale dependent thresholds.

Theorem 7 (Variational formulation of the wavelet–vaguelette estimator)**.** *For any* $r > 0$*, the wavelet–vaguelette soft-thresholding estimator* $\hat{f}^{(\mathrm{WV})}$ *is the unique solution of*

$$\begin{cases} \min\limits_{f\in\mathcal{F}} & \sum\limits_{\lambda\in\Lambda} \left|\langle \phi_\lambda, f\rangle\right|^r \\ \text{s.t.} & \max\limits_{\lambda\in\Lambda} \dfrac{\left|\langle v_\lambda, g - \boldsymbol{K}\,f\rangle\right|}{\kappa_\lambda q_\lambda} \le 1\,. \end{cases}$$

The same holds true, if we replace the objective function above by $\|f\|^2_{L^2(\Omega)}$*.*

Reasonable ways of adjusting $\hat{f}^{(\mathrm{WV})}$ to the case of discretely sampled data $Y_i = (\boldsymbol{S}_n \circ \boldsymbol{K} \circ \boldsymbol{E}_p \beta)_i + \varepsilon_i$, with $\beta = \boldsymbol{S}_p(f) = (f(x_j))_{j=1}^p$, as in model (II) are discussed in Donoho (1995b, Section 6.3). As a consequence, $\hat{f}^{(\mathrm{WV})}$ can be viewed as a MIND estimator in (9). We omit details and restrict our representation to the continuous model (unlike the rest of the paper) for simplicity.

In Donoho (1995b) wavelet–vaguelette decompositions have been derived for integration, fractional integration and the Radon transform. Related techniques for the solution of statistical inverse problems can be found in Abramovich and Silverman (1998), Candès and Donoho (2002), and Kalifa and Mallat (2003). In Abramovich and Silverman (1998) the roles of wavelets and vaguelettes are reversed. In Kalifa and Mallat (2003) special mirror wavelet bases are constructed for certain deconvolution problems. The use of curvelets instead of wavelets was studied in Candès and Donoho (2002) and applied to the noisy Radon inversion.

We stress that the existence of a wavelet–vaguelette decomposition requires a certain scale invariance of the operator $\boldsymbol{K}$, which are not satisfied in general, see Donoho (1995b).

4 Variational methods for estimation

In the last section we studied thresholding approaches using explicit expansions with respect to (multiscale) systems of functions (dictionaries). We derived equivalent variational formulations as constrained optimization problems with a multiscale constraint. The thresholding based multiscale approaches can be combined with variational regularization where no explicit dictionary is given. Again, they are casted as instances of the MIND in (9).

In this section we provide a selective overview of such variational estimation schemes. We start with standard penalized least squares and then present hybrid approaches combining variational and multiresolution schemes.

4.1 Penalized least squares

Presumably the most basic and prominent variational estimation techniques is *penalized least squares*

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|Y - \boldsymbol{X}\beta\|_2^2 + \gamma R(\beta) \,.$$

Here $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is some regularization functional and $\gamma > 0$ a penalty parameter. In the special case $R(\beta) = \|\beta\|_2^2$ the penalized least squares is known as *ridge regression*.

By interpreting $1/\gamma > 0$ as Lagrange multiplier, the penalized least squares estimator can be written in the constrained form

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^p} & R(\beta) \\ \text{s.t.} & \|Y - \boldsymbol{X}\beta\|_2 \leq q \,. \end{cases}$$

In fact, the constrained optimization and its unconstrained version are essentially equivalent if the Lagrangian parameter is chosen according to Morozov's discrepancy principle (see Theorem 14 in the Appendix for a precise statement). The constrained optimization problem is obviously a particular case of the MIND in (9) if one considers the identity $\boldsymbol{I}\colon \mathbb{R}^n \to \mathbb{R}^n$ as the only probe functional. In this sense, it has only a single scale.

Another well studied instance of regularized least squares are the lasso and its variants that will be studied in Section 4.3.

4.2 Total variation regularization

Thresholding based multiscale methods provide spatial adaptivity and are known to be optimal for function estimation in Besov balls. However, they often show visually disturbing artifacts due to missing band pass information. In contrast, variational methods, such as total variation regularization, often produce visually more appealing results.

Choosing the one-dimensional discrete total variation as regularization functional, the constrained least squares estimator takes the form

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^p} & \sum\limits_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \\ \text{s. t.} & \|Y - \boldsymbol{X}\beta\|_2 \le q \, . \end{cases} \tag{17}$$

This is *discrete* total variation regularization, also known as *trend filtering* (Tibshirani, 2014). In the one-dimensional nonparametric regression case (model (I)), a variant with *continuous* total variation as regularization instead is studied and shown to be minimax optimal over bounded variation function classes in Mammen and van de Geer (1997); See also Davies and Kovac (2001) and Davies et al. (2009). In the case that β represents an image, the two-dimensional analogs and the penalized version are instances of the Rudin–Osher–Fatemi (ROF) denoising model introduced in Rudin et al. (1992). There are extensions to higher orders of total variation and/or general dimensions (Hütter and Rigollet, 2016, Sadhanala et al., 2017, Fang et al., 2021, Guntuboyina et al., 2020, Ortelli and van de Geer, 2020), to tensors (Ortelli and van de Geer, 2021), and also to graphs (Wang et al., 2016). See also Chambolle and Lions (1997), where explicit relations between constraint and unconstraint total variation minimization have been derived in a very general (infinite dimensional) setting. All these approaches regularize the total variation functional in a global "monoscale" fashion, as in (17). In the following, we will discuss its multiscale extension to systems of scales.

This is motivated from empirical studies which show that monoscale total variation methods seem to lack spatial adaptivity to varying smoothness of the underlying signal or image. See e.g. the discussion in Candès and Guo (2002, Section 5.1), and also Section 6.

Hence, one wishes to combine "the best of both worlds", by solving

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^p} & \|\beta\|_{\mathrm{TV}} \\ \text{s. t.} & \max\limits_{\lambda \in \Lambda} \dfrac{|\langle \phi_\lambda, Y - \boldsymbol{X}\beta \rangle|}{q_\lambda} \le 1 \, , \end{cases} \tag{18}$$

where $\|\cdot\|_{TV}$ represents the total variation, and $(\phi_\lambda)_{\lambda \in \Lambda}$ is a multiscale system such as a wavelet basis or a shearlet frame.

The hybrid approach of (18) has been introduced independently by several authors. In Candès and Guo (2002) and Starck et al. (2011) the formulation in (18) is proposed for $\boldsymbol{X} = \boldsymbol{I}$ in combination with overcomplete dictionaries. Chan and Zhou (2000) and Durand and Froment (2001) used a wavelet basis. Finally, Malgouyres (2002a,b) studied (18) for general $\boldsymbol{X}$. In del Álamo et al. (2021), the hybrid total variation estimator for $\boldsymbol{X} = \boldsymbol{I}$ in (18) with various dictionaries, e.g. wavelets, curvelets or shearlets, is shown to be asymptotically minimax optimal (up to a logarithmic factor) with respect to L^q-risk ($1 \le q < \infty$) for the estimation of bounded variation functions on $[0, 1]^d$ with $d \ge 1$; Such rates in L^2-risk for the bounded variation space match those for the inscribed Sobolev space $W^{1,1}$ when $d \le 2$, but turn out to be slower for $d \ge 3$ (a phase transition). Similar statistical justifications of the estimator in (18) have been established for linear inverse problems in del Álamo and Munk (2020).

4.3 The (group) lasso and Nemirovskii's estimator

Let $(\phi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal basis of $\mathbb{R}^p$. Suppose that the index set Λ is written as

$$\Lambda = \bigcup_{a\in\mathcal{A}} \Lambda_a \,,$$

with possible overlapping (though the later theory assumes disjointness) subsets $\Lambda_a \subset \Lambda$. We further define, for every $a \in \mathcal{A}$, the linear mapping $\mathbf{\Phi}_a \colon \mathbb{R}^p \to \mathbb{R}^{\Lambda_a}$,

$$\mathbf{\Phi}_a\,\beta := (\langle \phi_\lambda, \beta\rangle : \lambda \in \Lambda_a) \qquad \text{for } \beta \in \mathbb{R}^p \,.$$

The *adaptive group lasso* is then defined via the penalized optimization (cf. (4))

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\,\|Y - \boldsymbol{X}\beta\|_2^2 + \gamma \sum_{a\in\mathcal{A}} w_a \|\mathbf{\Phi}_a\,\beta\|_2 \,. \tag{19}$$

The following theorem (based on Fenchel duality) shows that the adaptive group lasso estimator is also a special instance of the MIND in (9).

Theorem 8 (Dual formulation of the adaptive group lasso)**.** *Suppose that* $\Lambda = \bigcup_{a\in\mathcal{A}} \Lambda_a$ *consists of disjoint subsets* Λ_a. *Then the adaptive group lasso in* (19) *and the constrained optimization problem*

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^p} & \dfrac{1}{2}\,\|\boldsymbol{X}\beta\|_2^2 \\ \text{s.t.} & \max\limits_{a\in\mathcal{A}} \dfrac{\left\|\mathbf{\Phi}_a\,\boldsymbol{X}^\mathsf{T}(Y - \boldsymbol{X}\beta)\right\|_2}{w_a} \le \gamma \,, \end{cases} \tag{20}$$

have the same sets of solutions.

In the special case that each subset Λ_a consists of a single element, the adaptive group lasso estimator equals the adaptive lasso estimator (Zou, 2006, Huang et al., 2008). If additionally the weights are taken equal to one, then one obtains the standard lasso estimator

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\,\|Y - \boldsymbol{X}\beta\|_2^2 + \gamma \sum_{j=1}^{p} |\beta_j| \,,$$

which is introduced in Tibshirani (1996).

The following special case of Theorem 8 is well known, also in the compressed sensing and sparse recovery community (Fuchs, 2001, 2004, Tropp, 2006). A related result has been obtained in Osborne et al. (2000) for the lasso with the ℓ^1-norm as constraint.

Corollary 9 (Dual formulation of the lasso)**.** *The standard lasso estimator and the constrained optimization problem*

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^p} & \dfrac{1}{2}\,\|\boldsymbol{X}\beta\|_2^2 \\ \text{s.t.} & \left\|\boldsymbol{X}^\mathsf{T}(Y - \boldsymbol{X}\beta)\right\|_\infty \le \gamma \,, \end{cases}$$

have the same sets of solutions.

The dual characterization of the lasso in Corollary 9 reveals the close connection between the Dantzig selector and the lasso: Both estimators use the same constraint $\left\|\boldsymbol{X}^{\mathsf{T}}(Y-\boldsymbol{X}\beta)\right\|_{\infty} \leq \gamma$. This connection has been exploited in Bickel et al. (2009). However, among all feasible elements the Dantzig selector minimizes the ℓ^1-norm of β, whereas the lasso minimizes the ℓ^2-norm of the prediction $\boldsymbol{X}\beta$. Hence, in general, the lasso is better for prediction, while the Dantzig selector is better for coefficient estimation.

As a further consequence of Theorem 8, we can show that in the case of $\boldsymbol{X} = \boldsymbol{I}$ the adaptive group lasso estimator is equal to the block soft-thresholding.

Corollary 10 (Equivalence of the block thresholding and the adaptive group lasso)**.** *Consider the regression case $\boldsymbol{X} = \boldsymbol{I}$ and suppose that $\Lambda = \bigcup_{a\in\mathcal{A}} \Lambda_a$ consists of disjoint subsets Λ_a. Then the following three estimators coincide:*

(a) *The block soft-thresholding estimator in* (15)*.*

(b) *The adaptive group lasso estimator in* (19)*.*

(c) *The multiscale Dantzig estimator in* (20)*.*

Finally, we stress that in exactly the same way Nemirovskii's estimator can be shown to be equivalent to its reverse constrained variant.

Theorem 11 (Reverse formulation of Nemirovskii's estimator)**.** *Nemirovskii's estimator in* (8) *is equivalent to the reverse formulation*

$$\begin{cases} \min\limits_{\beta\in\mathbb{R}^n} \|\beta\|_{k,q} \\ \text{s.t.}\ \|Y-\beta\|_{\mathcal{N}} \leq q, \end{cases}$$

for some proper choice of q.

Thus, Nemirovskii's estimator is also an instance of MIND in (9). The reverse version was indeed introduced by Nemirovskii (1985), who credited the original idea to S. V. Shil'man. It was shown to be adaptively minimax optimal over Sobolev balls in Grasmair et al. (2018). The reverse Nemirovskii's estimator appears to us favorable over the original and penalized version, as the threshold q has a distinct statistical interpretation (Section 2).

4.4 Multiscale change point segmentation

As a special case, we consider now the model (III) of change point detection in detail. The target is to estimate a piecewise constant function $f:[0,1]\to\mathbb{R}$ with values $f(x_i)=\sum_{k=1}^{i}\beta_k$ for $i=1,\ldots,n$. A typical choice of regularity measure is the number of jumps, that is, $R(\beta)=\|\beta\|_0$, which is defined as the number of non-zero elements of β. The corresponding *jump penalized least squares* estimator (Boysen et al., 2009) takes the form of

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|Y-\boldsymbol{X}\beta\|_2^2 + \gamma\|\beta\|_0 .$$

Refined penalties can be found in e.g. Zhang and Siegmund (2007) and Davis and Yau (2013).

In practice, the selection of the global penalty parameter γ is tricky for jump penalized least squares, in particular, when the change points are spatially inhomogeneous. Frick et al. (2014)

introduced a remedy, SMUCE (simultaneous multiscale change point estimator) following the MIND idea, by combining variational estimation with multiple tests on residuals over different scales. More generally, *multiscale change point segmentation* (MCPS; Li et al., 2019) is defined as any solution to the constrained non-convex optimization problem

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^n} \quad \|\beta\|_0 \\ \text{s.t.} \quad \max\limits_{(i,j)\in\mathcal{I}_\beta} \left\{ \dfrac{\left|\sum_{k=i}^{j}(Y - \boldsymbol{X}\beta)_k\right|}{\sqrt{j-i+1}} - s_{i,j} \right\} \le q\,, \end{cases} \tag{21}$$

where $\mathcal{I}_\beta = \{(i,j) : 1 \le i \le j \le n, \text{ and } \beta_k = 0 \text{ for } i < k \le j\}$, and $s_{i,j}$ are certain scale penalties, for instance, $s_{i,j} = \sqrt{2\log(n/(j-i+1))}$. In particular, SMUCE and its FDR variant, FDRSeg (Li et al., 2016), are instances of the MCPS. It can be easily seen that every MCPS is in fact a special instance of the MIND in (9). Similar to the general strategy for MIND, the threshold q can be set as the $(1-\alpha)$-quantile of

$$T_n = \max_{1\le i\le j\le n} \left\{ \frac{\left|\sum_{k=i}^{j} \varepsilon_k\right|}{\sqrt{j-i+1}} - s_{i,j} \right\}, \tag{22}$$

which, under no assumption, guarantees uniformly over β and $f = \boldsymbol{X}\beta$ that

$$\mathbf{P}_\beta\left\{\|\hat\beta\|_0 \le \|\beta\|\right\} = \mathbf{P}_f\left\{\#\text{jumps of } \hat f_n \ \le\ \#\text{jumps of } f_n\right\} \ge 1-\alpha\,,$$

where $\hat\beta$ is computed by the MCPS, $\hat f_n = \boldsymbol{X}\hat\beta$ and $f_n = \boldsymbol{X}\beta$ (recall (11) in Section 2). A comprehensive discussion of statistical optimality properties are provided in Frick et al. (2014), Li et al. (2016) and Li et al. (2019). Extensions can be found e.g. in Pein et al. (2017) for heterogeneous Gaussian error, in Vanegas et al. (2021) for general independent data, in Dette et al. (2020) for dependent data, and in Li et al. (2020) for automatic selecting the bins in a histogram and exploratory data analysis. Besides, Behr et al. (2018) extended the MCPS to blind source separation, more precisely, to recover piecewise constant functions (taking values in a finite set) from noisy measurements of their mixtures. In addition, estimators similar to MCPS, but with different regularizations accounting for shape or smoothness have been considered in Davies and Kovac (2001), Dümbgen and Spokoiny (2001), Davies and Meise (2008), Davies et al. (2009) and Schmidt-Hieber et al. (2013).

5 Distributional properties and selection of the threshold

As discussed in Section 2, the threshold q for MIND is fully determined by the quantiles of multiscale statistic T_n in (12), which can be estimated by Monte Carlo simulations. As an alternative and computationally more efficient approach, such quantiles can be computed via either limiting distributions of T_n or (approximate) tail probabilities and bounds of T_n. Here we focus on the first approach, while for the latter we refer to e.g. Siegmund and Yakir (2000), Fang et al. (2020) and the references therein.

5.1 Case of orthogonal bases

Suppose for the moment that $(\phi_{n,\lambda})_{\lambda\in\Lambda_n}$ is an orthonormal basis of $\mathbb{R}^n$. A particular important class of probe functionals takes $\mathbf{\Pi}_{n,\lambda}\,\varepsilon_n := \langle\phi_{n,\lambda},\varepsilon_n\rangle$. Under the i.i.d. Gaussian assumption, these functionals are i.i.d. Gaussian again, and hence

$$\lim_{n\to\infty}\mathbf{P}\Big\{\max_{\lambda\in\Lambda_n}|\langle\phi_{n,\lambda},\varepsilon_n\rangle| \;\le\; \sigma\sqrt{2\log n}\;+\;\sigma\,\frac{2x-\log\log n-\log\pi}{2\sqrt{2\log n}}\Big\} \;=\; \exp\left(-e^{-x}\right)\,, \quad (23)$$

where the limit distribution is known as the *Gumbel extreme value distribution.* In the first order, q equals $\sigma\sqrt{2\log n}$, which corresponds to the asymptotic behavior of the maximum of absolute values of i.i.d. Gaussian random variables, and is the universal threshold as proposed in the seminal work of Donoho (1995a), recall Section 3.1.

5.2 Redundant systems

Now suppose that $(\phi_{n,\lambda})_{\lambda\in\Lambda_n}$ is a redundant frame instead of an orthonormal basis and consider again the probe functionals $\mathbf{\Pi}_{n,\lambda}\,\varepsilon_n := \langle\phi_{n,\lambda},\varepsilon_n\rangle$. In this situation finding the distribution of the multiscale statistic T_n in (12) is more involved than in the independent case. In Haltmeier and Munk (2014) similar asymptotic distributions as in (23) have been derived for a wide class of redundant systems:

Definition 12 (Asymptotically stable frames)**.** *For any $n\in\mathbb{N}$, let $\mathcal{D}_n := (\phi_{n,\lambda})_{\lambda\in\Lambda_n}$ be a frame of $\mathbb{R}^n$ with upper frame bound b_n, as in* (16)*. Then $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ is called an* asymptotically stable *family of frames, if $\|\phi_{n,\lambda}\|_2 = 1$ for all $n\in\mathbb{N}$ and all $\lambda\in\Lambda_n$, $\sup\{b_n : n\in\mathbb{N}\} < \infty$, and $|\{(\lambda,\mu) : |\langle\phi_{n,\lambda},\phi_{n,\mu}\rangle| \ge \rho\}| = o\big(|\Lambda_n|/\sqrt{\log|\Lambda_n|}\big)$ for some $\rho<1$.*

Roughly speaking, for such frames, correlations of $\langle\phi_{n,\lambda},\varepsilon_n\rangle$ asymptotically vanish fast enough and hence the system $(\phi_{n,\lambda})_{\lambda\in\Lambda_n}$ asymptotically behaves as an orthonormal system. Many frames used in applications, such as unions of bases, non-redundant and redundant wavelet systems, and curvelet frames are covered by Definition 12 (Haltmeier and Munk, 2014). Hence, this justifies the universal thresholding by $\sigma\sqrt{2\log|\Lambda_n|}$ for many systems beyond wavelets.

In case of one dimension and the probe functionals as indicators of all subintervals, the multiscale statistic T_n takes the particular form of (22). Here the probe functionals are a strongly redundant frame, and thus not asymptotically stable. However, similar asymptotic distributional results as in (23) still hold, see Siegmund and Venkatraman (1995) and Siegmund and Yakir (2000). Generalization to higher dimension can be found in Kabluchko (2011), Proksch et al. (2018) and König et al. (2020).

6 Numerical computation

In the previous sections we have seen that many estimation techniques can be written as instances of the MIND in (9). For thresholding methods, the solution of (9) is often given by an explicit formula (see Section 3). In the general case, however, a MIND must be computed numerically by applying an optimization procedure.

In case that the regularization functional R is convex, (9) is a nonsmooth convex optimization, and often of large size (when there are many probe functionals). Recent development (e.g.

Nesterov, 2005, Beck and Teboulle, 2009, Becker et al., 2011, Chambolle and Pock, 2011) in optimization has made the computation of MIND in (9) feasible (e.g. in a few minutes for 256×256 images) on standard laptops, see e.g. Frick et al. (2012). Run time comparisons suggest the primal-dual hybrid-gradient algorithm (Chambolle and Pock, 2011) as a powerful general computational scheme for MIND, see del Álamo et al. (2020) for details and a comparison to other optimization methods, including semismooth Newton and ADMM (alternating direction method of multipliers). MATLAB codes are freely available at `https://github.com/housenli/MIND`.

In case of nonconvex R, the same algorithm can be applied but there is no guarantee for global optimality, in general. However, in the particular case of change point segmentation (see Section 4.4), the global optimal solutions for MCPS in (21) can be computed using dynamic programming algorithms together with speedups leading to run time of the order $O(n \log n)$ in most often cases. Implementation is made available in R packages (e.g. `stepR` or `FDRSeg`) on CRAN, see Frick et al. (2014), Li et al. (2016) and Pein et al. (2017) for details. Besides, approximate solutions can be found in sublinear run time (Kovács et al., 2020).

6.1 Image denoising

We consider an example of model (I), where $\beta \in \mathbb{R}^n$ is an $m \times m$ image with $n = m^2$, often known as *image denoising*. We numerically compare several instances of MIND: wavelet soft-thresholding (Section 3.1), total variation penalized least squares (Section 4.1), and two hybrid approaches (Section 4.2) combining total variation with wavelets and shearlets, respectively. Daubechies' symlets with six vanishing moments and shearlets with four scale levels (default in Kutyniok et al., 2012) are used. The threshold for wavelet soft-thresholding and hybrid methods is set as the 90%-quantile of the corresponding multiscale statistic in (12), respectively. The penalty parameter γ in total variation penalized least squares is tuned to give the best visual quality (cf. Section 1.2.1).

Results are depicted in **Figure 3**. In this example, wavelet soft-thresholding performs the worst; In particular, it shows artificial oscillations across discontinuity (common for wavelets due to missing band pass information). The total variation as a regularization functional has a comparably better performance, while blurring out several details (e.g. at the bottom right conner). The hybrid combination with wavelets leads to an improved performance, because more than one spatial scale is incorporated (recall Section 4.2). The hybrid method of total variation and shearlets produces clearly the best result as it recovers features over a range of spatial scales quite accurately. It seems that visual inspection of the image is compatible with the PSNR (peak signal-to-noise ratio).

6.2 Detection of change points

Recall example (III). On a randomly generated piecewise constant signal, we run the jump penalized least squares, and two multiscale methods, SMUCE (Frick et al., 2014) and FDRSeg (Li et al., 2016), all of which are instances of the MIND (Section 4.4). The penalty parameter $\gamma = 2\sigma^2 \log_2 n$, which corresponds to the Bayesian information criterion, is used for the jump penalized least squares. For SMUCE and FDRSeg, the default parameters in R packages `stepR` and `FDRSeg`, respectively, are used. The comparison is shown in **Figure 4**. The jump penalized least squares recovers major structure of the signal, but misses three jumps. SMUCE examines over multiple scales (here lengths of intervals) and recovers one more jump. By switching to a

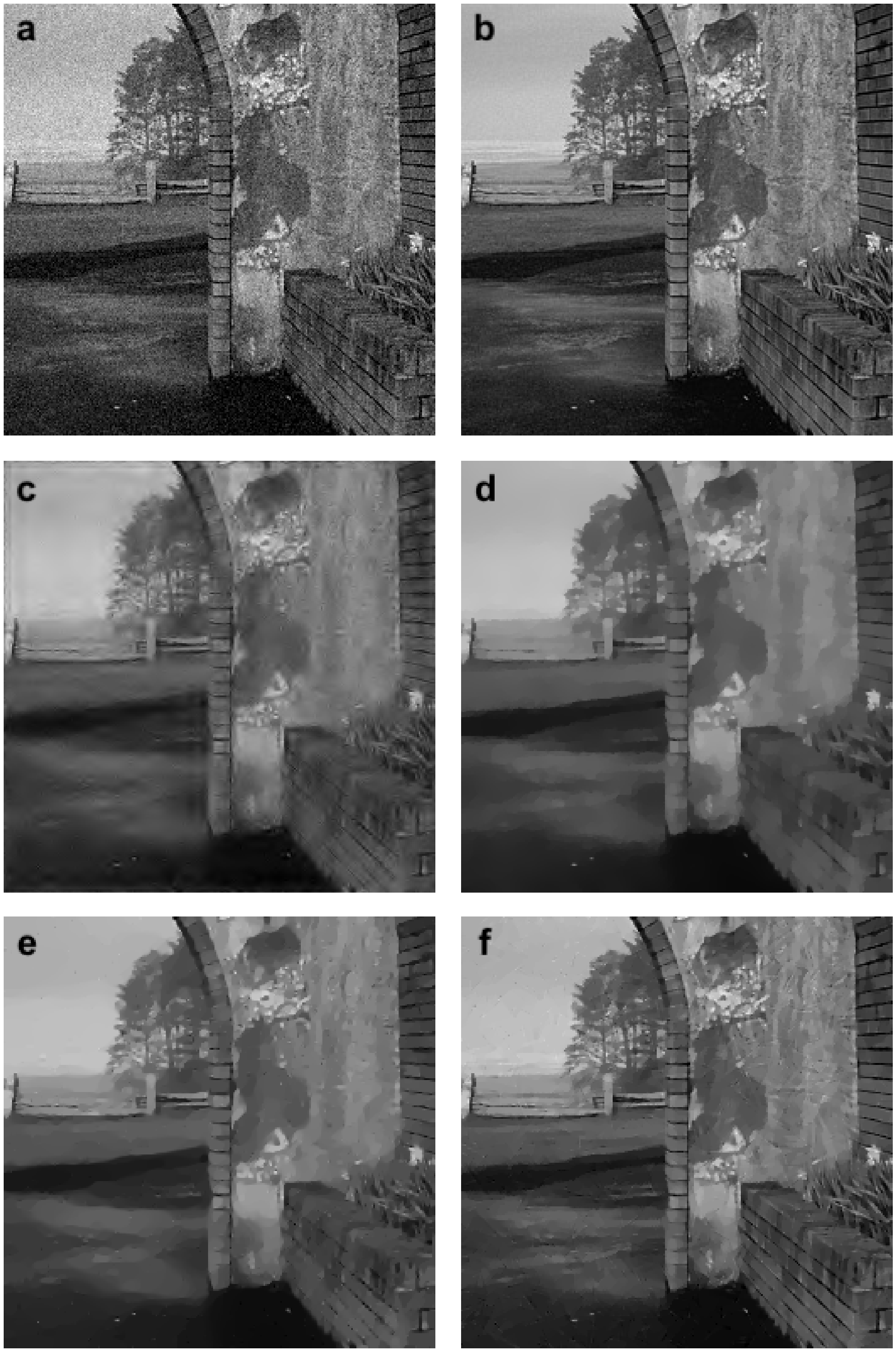

Figure 3: Image denoising via different MINDs. (a) Noisy data with $\sigma = 14$ (PSNR = 25.2); (b) True image from BSDS500 (Martin et al., 2001); (c) Wavelet soft-thresholding (PSNR = 25.4); (d) Total variation penalized least squares (ROF; PSNR = 25.9); (e) Hybrid wavelet–total variation (PSNR = 26.1); (f) Hybrid shearlet–total variation (PSNR = 28.2).

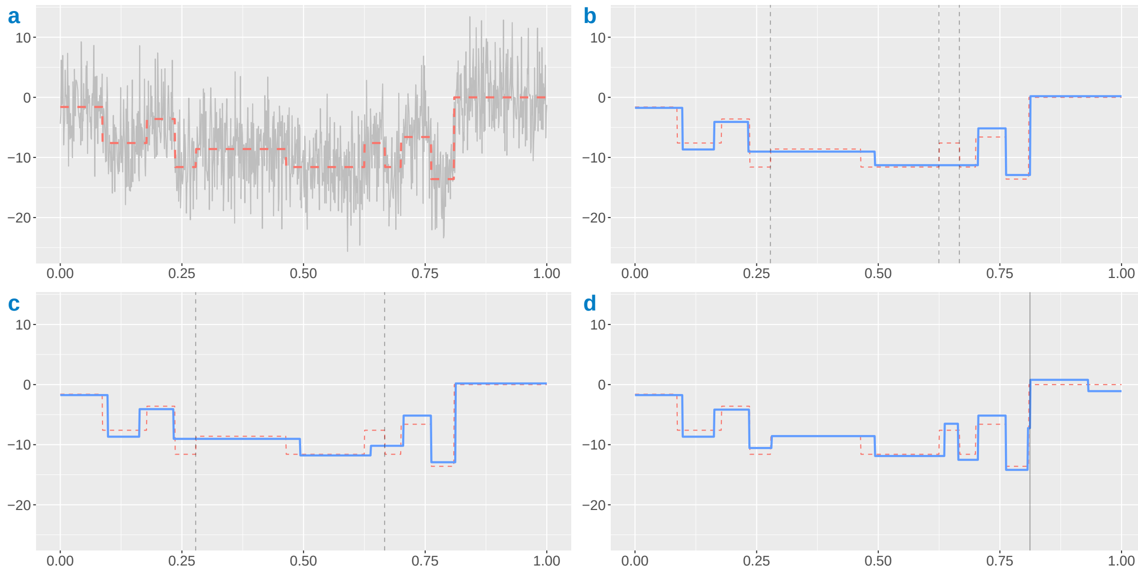

Figure 4: Change point estimation for a randomly generated signal. (a) Noisy data (solid gray line); (b)–(d) Estimators (solid blue line) of jump penalized least squares, SMUCE, and FDRSeg, respectively. Vertical dashed lines indicate missed change points (false negatives), while vertical solid lines mark artificial change points (false positives). The true signal (dashed red line) is plotted in all panels.

weaker error criterion, FDRSeg detects all jumps of the signal, but at the price of including one artificial jump.

Acknowledgement

We thank Shayan Hundrieser, Russell Luke, Marc Teboulle, Frank Werner for helpful comments and in particular Timo Aspelmeier for inspiring discussions and computational assistance.

A Relations between constrained and unconstrained minimization

There are at least two close connections between the constrained optimization in (5) and the unconstrained optimization in (4). The first one arises from a Lagrangian multiplier approach, whereas the second arises from Fenchel duality. In what follows we analyze both approaches.

A.1 Lagrangian multiplier approach

Let $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $G\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex proper functionals, and let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. It is convenient to allow a convex functional to attain the value ∞. The set $\mathcal{D}(R) := \{\beta \in \mathbb{R}^p\colon R(\beta) < \infty\}$ where the functional R takes finite values is referred to as the *domain* of R. The functional R is called *proper*, if $\mathcal{D}(R) \neq \emptyset$, see Rockafellar (1970).

We will investigate the relation between the following optimization problems:

$$\min\{R(\beta)\colon G(\boldsymbol{X}\beta) \leq q\} \qquad (G\text{-constrained}) \qquad (24)$$

$$\min\{G(\boldsymbol{X}\beta) + \gamma R(\beta)\colon \beta \in \mathbb{R}^p\} \qquad (\text{penalized}) \qquad (25)$$

$$\min\{G(\boldsymbol{X}\beta)\,:\, R(\beta) \leq c\} \qquad (R\text{-constrained}), \qquad (26)$$

for parameters q, γ, $c \geq 0$. Here we simplify the notation by $G(\boldsymbol{X}\beta) \equiv G(\boldsymbol{X}\beta; Y)$. We will show that (24), (25), and (26) are equivalent in the sense that any solution $\beta_\star$ of one of these optimization problems also provides a solution of the others.

The proof will use the following saddle-point theorem, which is a well known result from convex analysis.

Lemma 13. *Suppose that $\Phi, \Psi\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ are proper convex functionals and denote by $\mathcal{L}\colon \mathbb{R}^p \times \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$ the Lagrangian defined by $\mathcal{L}(\beta, \mu) = \Phi(\beta) + \mu\Psi(\beta)$.*

(a) *Let $(\beta_\star, \mu_\star) \in \mathbb{R}^p \times \mathbb{R}_{\geq 0}$ be a* saddle-point *of $\mathcal{L}$, that is, for all $(\beta, \mu) \in \mathbb{R}^p \times \mathbb{R}_{\geq 0}$,*

$$\mathcal{L}(\beta_\star, \mu) \leq \mathcal{L}(\beta_\star, \mu_\star) \leq \mathcal{L}(\beta, \mu_\star)\,.$$

Then $\beta_\star \in \operatorname{argmin}\{\Phi(\beta) : \Psi(\beta) \leq 0\}$.

(b) *If, additionally, Slater's condition*

$$\mathcal{D}(\Phi) \cap \{\beta \in \mathbb{R}^p : \Psi(\beta) < 0\} \neq \emptyset$$

holds and $\beta_\star \in \operatorname{argmin}\{\Phi(\beta) : \Psi(\beta) \leq 0\}$, then there exists a Lagrangian multiplier $\mu_\star \geq 0$ such that $(\beta_\star, \mu_\star)$ is a saddle-point of $\mathcal{L}$.

Proof. Part (a) follows from Theorem 28.3 in Rockafellar (1970).

Part (b) follows from Corollary 28.2.1 and Theorem 28.3 in Rockafellar (1970). □

We next derive the equivalence of the G-constrained and penalized formulation, i.e. of (24) and (25).

Theorem 14 (Equivalence of (24) and (25))**.** *Let $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $G\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper convex functionals and let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Then* (24) *and* (25) *are equivalent in the following sense:*

(a) *If $\beta_\star$ is a solution of* (25) *for some $\gamma > 0$, then $\beta_\star$ solves* (24) *with $q = G(\boldsymbol{X}\beta_\star)$.*

(b) *For a given $q > 0$, suppose that $G(\boldsymbol{X}\beta_0) < q$ for some $\beta_0 \in \mathcal{D}(R)$ and that we have*

$$\operatorname{argmin}(R) \cap \{\beta\colon G(\boldsymbol{X}\beta) \leq q\} = \emptyset\,. \qquad (27)$$

Then, any solution $\beta_\star$ of (24) *satisfies $G(\boldsymbol{X}\beta_\star) = q$ and there exists some $\gamma > 0$ such that $\beta_\star$ also solves* (25).

In the special case of strictly convex penalties R and $G(\boldsymbol{X}\beta) = \frac{1}{2}\|Y - \boldsymbol{X}\beta\|^2$, Theorem 14 seems to be well known but it is hard to find an accessible reference. For an infinite dimensional setting such a proof can be found in Vasin (1970) and Ivanov et al. (2002). Below we include a short proof for the convex penalties (not necessarily strict convex), but for finite dimensional ground space. Note that if R is strictly convex and z_0 denotes its unique minimizer, then the condition of (27) reads $G(\boldsymbol{X}z_0) > q$, which simply means that the unique minimizer of R is not feasible for (24).

Proof of Theorem 14. Part (a): Suppose first that $\beta_\star$ is a solution of (25) for some $\gamma > 0$ and take $q = G(\boldsymbol{X}\beta_\star)$, $\Psi(\beta) = G(\boldsymbol{X}\beta) - q$. Then we have $\Psi(\beta_\star) = 0$ which implies that $R\big(\beta_\star\big) + \mu\Psi\big(\beta_\star\big) = R\big(\beta_\star\big) + \frac{1}{\gamma}\Psi\big(\beta_\star\big)$ for all $\mu \geq 0$. Further, since $\beta_\star$ is a minimizer of $G(\boldsymbol{X}\beta; Y) + \gamma R(\beta)$, we also have $\gamma R(\beta_\star) + (G(\boldsymbol{X}\beta_\star) - q) \leq \gamma R(\beta) + (G(\boldsymbol{X}\beta) - q)$, which yields $R(\beta_\star) + \frac{1}{\gamma}\Psi(\beta_\star) \leq R(\beta) + \frac{1}{\gamma}\Psi(\beta)$ for all $\beta \in \mathbb{R}^p$. In total, we have verified that

$$R\big(\beta_\star\big) + \mu\Psi\big(\beta_\star\big) \leq R\big(\beta_\star\big) + \frac{1}{\gamma}\Psi\big(\beta_\star\big) \leq R(\beta) + \frac{1}{\gamma}\Psi(\beta) \qquad \text{all } (\beta, \mu) \in \mathbb{R}^p \times \mathbb{R}_{\geq 0}\,.$$

Hence $(\beta_\star, 1/\gamma)$ is a saddle-point of the Lagrangian $\mathcal{L}(\beta, \mu) = R(\beta) + \mu\Psi(\beta)$. Lemma 13 implies that $\beta_\star \in \operatorname{argmin}\{R(\beta) \colon \Psi(\beta) \leq 0\}$. Recalling that $\Psi(\beta) = G(\boldsymbol{X}\beta) - q$ shows that $\beta_\star$ is a solution of (24).

Part (b): Now suppose that $\beta_\star \in \mathbb{R}^p$ is a solution of the constrained problem in (24) for some $q > 0$. According to our assumption there exists $\beta_0 \in \mathbb{R}^p$ with $G(\boldsymbol{X}\beta_0) < q$. Consequently, the convex functionals $R(\beta)$ and $\Psi(\beta) = G(\boldsymbol{X}\beta) - q$ satisfy Slater's condition. According to Lemma 13 the corresponding Lagrangian $\mathcal{L}(\beta, \mu) = R(\beta) + \mu\Psi(\beta)$ admits a saddle-point $(\beta_\star, \mu_\star) \in \mathbb{R}^p \times \mathbb{R}_{\geq 0}$. In particular, $\beta_\star$ minimises $R(\beta) + \mu_\star G(\boldsymbol{X}\beta)$. It remains to show that $\mu_\star > 0$. To that end, suppose to the contrary that $\mu_\star = 0$. Then $\beta_\star \in \operatorname{argmin} R$. Since $\beta_\star$ is also a solution of (24), it in particular satisfies the constraint in (24), which implies

$$\beta_\star \in \operatorname{argmin}(R) \cap \{\beta \in \mathbb{R}^p \colon G(\boldsymbol{X}\beta) \leq q\}$$

This contradicts the assumption in (27). Hence $\mu_\star > 0$ and further $G(\boldsymbol{X}\beta_\star) = q$. This concludes the proof of the theorem after taking $\gamma = 1/\mu_\star$. □

Next we show the equivalence between the penalized and R-constrained formulation, i.e. (25) and (26). Together with Theorem 14 this also implies that the G- and R-constrained formulation are equivalent, too.

Theorem 15 (Equivalence of (25) and (26)). *Let $R \colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $G \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper convex functionals and let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Then* (25) *and* (26) *are equivalent in the following sense:*

(a) *If $\beta_\star$ is a solution of* (25) *for some $\gamma > 0$, then $\beta_\star$ solves* (26) *with $c = R(\beta)$.*

(b) *For a given $c > 0$, suppose $R(\beta_0) < c$ and $G(\boldsymbol{X}\beta_0) < \infty$ for some $\beta_0 \in \mathbb{R}^p$ and let $\beta_\star$ be a solution of* (26)*. Then there exists $\gamma \geq 0$ such that $\beta_\star$ solves* (25) *and $\gamma(R(\beta_\star) - c) = 0$.*

Proof. Part (a): Suppose first that $\beta_\star$ is a solution of (25) for some $\gamma > 0$ and take $c = R(\boldsymbol{X}\beta_\star)$, $\Phi(\beta) = G(\boldsymbol{X}\beta)$, $\Psi(\beta) = R(\beta) - c$. Then we have $\Psi(\beta_\star) = 0$ which implies that $\Phi(\beta_\star) + \mu\Psi(\beta_\star) = \Phi\big(\beta_\star\big) + \gamma\Psi(\beta_\star)$ for all $\mu \geq 0$. Further, since $\beta_\star$ is a minimizer of $G(\boldsymbol{X}\beta) + \gamma R(\beta)$, we also have $G(\boldsymbol{X}\beta_\star) + \gamma(R(\beta_\star) - c) \leq G(\boldsymbol{X}\beta) + \gamma(R(\beta) - c)$ for $\beta \in \mathbb{R}^p$, which yields $\Phi(\beta_\star) + \gamma\Psi(\beta_\star) \leq \Phi(\beta) + \gamma\Psi(\beta)$. In total, we have verified the inequalities

$$\Phi(\beta_\star) + \mu\Psi(\beta_\star) \leq \Phi(\beta_\star) + \gamma\Psi(\beta_\star) \leq \Phi(\beta) + \gamma\Psi(\beta) \quad \text{ for } (\beta, \mu) \in \mathbb{R}^p \times \mathbb{R}_{\geq 0}\,.$$

Hence $(\beta_\star, \gamma)$ is a saddle-point of the corresponding Lagrangian $\mathcal{L}(\beta, \mu) = \Phi(\beta) + \mu\Psi(\beta)$. Lemma 13 therefore implies that $\beta_\star \in \operatorname{argmin}\{\Phi(\beta) : \Psi(\beta) \leq 0\}$. Recalling that $\Phi(\beta) = G(\boldsymbol{X}\beta)$, $\Psi(\beta) = R(\beta) - c$ shows that $\beta_\star$ is a solution of (24).

Part (b): Suppose that $\beta_\star \in \mathbb{R}^p$ is a solution of (26) of some $c > 0$. According to our assumption there exists $\beta_0 \in \mathbb{R}^p$ with $R(\beta_0) < q$. Consequently, the convex functionals $\Phi(\beta) = G(\boldsymbol{X}\beta)$ and $\Psi(\beta) = R(\beta) - c$ satisfy Slater's condition. According to Lemma 13 the corresponding Lagrangian $\mathcal{L}(\beta, \mu) = G(\boldsymbol{X}\beta) + \mu(R(\beta) - c)$ admits a saddle-point $(\beta_\star, \gamma) \in \mathbb{R}^p \times \mathbb{R}_{\geq 0}$. In particular, $\beta_\star$ minimises $G(\boldsymbol{X}\beta) + \gamma R(\beta)$. Since $\beta_\star$ is feasible, we have $R(\beta_\star) - c \leq 0$. Further, for all $\mu \geq 0$,

$$G(\boldsymbol{X}\beta_\star) + \mu(R(\beta_\star) - c) \leq G(\boldsymbol{X}\beta_\star) + \gamma(R(\beta_\star) - c),$$

which implies $\gamma(R(\beta_\star) - c) = 0$. □

We emphasize again that the relation between the threshold q and the penalty γ in Theorem 14 is only given in an implicit manner. Therefore, the unconstrained problem in (25) cannot be immediately used for solving the (more difficult) constrained problem in (24). The same statement applies to the equivalence of the penalized and R-constrained formulation. As shown in the next subsection, a more explicit relation between the G-constrained formulation and yet another different unconstrained optimization problem can be derived from the Fenchel duality.

The derived equivalences in particular apply to the MIND in (9) by taking

$$G(v) = \max_{a \in \mathcal{A}} \left\{ \frac{\|\boldsymbol{\Pi}_a (v - Y)\|_2}{w_a} - s_a \right\}$$

for $v \in \mathbb{R}^n$, where $\boldsymbol{\Pi}_a$ are given probe functionals, w_a and s_a certain weights and Y the given data, see Definition 1. Note that the functional G is obviously convex and proper.

A.2 Fenchel duality

For a proper convex functional $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$, we denote by

$$\begin{aligned} R^*\colon \mathbb{R}^p &\to \mathbb{R} \cup \{\infty\} \\ \mu &\mapsto \sup\{\langle \mu, \beta \rangle - R(\beta) : \beta \in \mathbb{R}^p\} \end{aligned}$$

the *Fenchel conjugate* of R, and by $\partial R(\beta_0)$ the *subdifferential* of R at β_0, i.e. $\beta^* \in \partial R(\beta_0)$ if and only if

$$R(\beta_0) < \infty \quad \text{and} \quad \langle \beta - \beta_0, \beta^* \rangle + R(\beta_0) \leq R(\beta) \ \text{ for all } \beta \in \mathbb{R}^p,$$

see Rockafellar (1970).

Definition 16. *Let $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $G\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be two proper convex functionals and let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Then,*

$$\min_{\beta \in \mathbb{R}^p} G(\boldsymbol{X}\beta) + R(\beta) \qquad \text{(primal)} \qquad (28)$$

$$\min_{\mu \in \mathbb{R}^n} G^*(\mu) + R^*(-\boldsymbol{X}^\top \mu) \qquad \text{(dual)} \qquad (29)$$

are referred to as the primal and dual minimization problem, respectively, corresponding to R, G and $\boldsymbol{X}$.

Theorem 17 (Fenchel's duality theorem)**.** *Suppose that* $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ *and* $G\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *are proper convex functionals, and let* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$*. Suppose further that there exists* $\beta_0 \in \mathbb{R}^p$ *such that* $R(\beta_0) < \infty$ *and that* G *is continuous at* $\boldsymbol{X}\beta_0$*. Then, for any* $(\beta_\star, \mu_\star) \in \mathbb{R}^p \times \mathbb{R}^n$*, the following statements are equivalent:*

(a) $\beta_\star$ *is a solution of the primal problem in* (28) *and* $\mu_\star$ *a solution of the dual problem in* (29)*.*

(b) *The Kuhn-Tucker conditions are satisfied, i.e.*

$$\mu_\star \in \partial G(\boldsymbol{X}\beta_\star) \tag{30}$$

$$-\boldsymbol{X}^\mathsf{T}\mu_\star \in \partial R(\beta_\star)\,. \tag{31}$$

Proof. See Rockafellar (1970, Section 31). □

If one takes $G(v) = \frac{1}{2}\|Y - v\|_2^2$, then Theorem 17 yields the following result:

Theorem 18. *Let* $R\colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ *be a proper convex functional and let* $Y \in \mathbb{R}^n$*. Then,*

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|Y - \boldsymbol{X}\beta\|_2^2 + R(\beta) \right\} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{X}\beta\|_2^2 + R^*(\boldsymbol{X}^\mathsf{T}(Y - \boldsymbol{X}\beta)) \right\}.$$

Proof. If $\beta_\star$ is a minimizer of $\frac{1}{2}\|Y - \boldsymbol{X}\beta\|_2^2 + R(\beta)$, then it is a solution of the primal problem in (28) corresponding to $G(v) = \frac{1}{2}\|Y - v\|_2^2$. In this case, one easily computes $G^*(\mu) = \frac{1}{2}\|\mu\|_2^2 + \langle Y, \mu\rangle$, see Rockafellar (1970). Now let $\mu_\star \in \mathbb{R}^n$ denote a solution of the dual problem in (29). Since R is proper there exists $\beta_0 \in \mathbb{R}^p$ such that $R(\beta_0) < \infty$ and obviously G is continuous at $\boldsymbol{X}\beta_0$. Hence we can apply Theorem 17 which implies that the Kuhn-Tucker conditions in (30) and (31) are satisfied. In particular, (30) implies $\mu_\star = \boldsymbol{X}\beta_\star - Y$, and therefore

$$\begin{aligned} G^*(\mu_\star) &= G^*(\boldsymbol{X}\beta_\star - Y) \\ &= \frac{1}{2}\|\boldsymbol{X}\beta_\star - Y\|_2^2 + \langle Y, \boldsymbol{X}\beta_\star - Y\rangle \\ &= \frac{1}{2}\|\boldsymbol{X}\beta_\star\|_2^2 - \frac{1}{2}\|Y\|_2^2\,. \end{aligned}$$

Because $\mu_\star$ is a solution of the dual problem in (29), this shows that $\beta_\star$ minimizes $\frac{1}{2}\|\boldsymbol{X}\beta\|_2^2 + R^*(\boldsymbol{X}^\mathsf{T}(Y - \boldsymbol{X}\beta))$. Similar arguments show that the converse relation also holds. □

B Proofs

Here we provide proofs for the assertions in the paper for the sake of completeness, noting that many of them are known but scattered in the literature. These assertions can be seen as special cases of the results obtained in Appendix A. But we favor simple and direct proofs whenever it is possible.

B.1 Proofs in Section 3

In this subsection, we give proofs for the assertions in Section 3 in the paper. Note that the following proofs do not rely on any results from Appendix A.

Proof of Theorem 3. Because $(\phi_\lambda)_{\lambda\in\Lambda}$ is an orthonormal basis, we can uniquely write any element in $\mathbb{R}^n$ as linear combination $\beta = \sum x_\lambda \phi_\lambda$ with coefficients $x_\lambda = \langle \phi_\lambda, \beta \rangle$. Hence $\hat\beta = \sum \hat x_\lambda \phi_\lambda$ is a solution of the stated optimization problem if and only if every coefficient $\hat x_\lambda$ is a solution of the one-dimensional optimization problem

$$\begin{cases} \min\limits_{x\in\mathbb{R}} & |x_\lambda|^r \\ \text{s.t.} & |\langle \phi_\lambda, Y\rangle - x_\lambda| \le q_\lambda \, . \end{cases}$$

The unique minimizer is given by the soft-thresholding $\hat x_\lambda = \eta^{(\text{soft})}\left(\langle \phi_\lambda, Y\rangle, q_\lambda\right)$, which yields the desired characterization using the objective $\sum_{\lambda\in\Lambda} |\langle \phi_\lambda, \beta\rangle|^r$ for any $r > 0$.

In the special case $r = 2$, we have $\sum_{\lambda\in\Lambda} |\langle \phi_\lambda, \beta\rangle|^2 = \sum_{i=1}^n \beta_i^2$, which shows the second claim. □

Proof of Theorem 4. This immediately follows from the variational characterization of soft-thresholding given in Theorem 3 and the relation

$$\eta^{(\text{JS})}(x, q) = \eta^{(\text{soft})}\left(x, \frac{q^2}{|x|}\right) \quad \text{for } x \neq 0 \, .$$

□

Proof of Theorem 5. The proof is elementary and similar to the one of Theorem 3. □

Proof of Theorem 7. By expanding any function $f \in L^2(\Omega)$ in a wavelet series $f = \sum_{\lambda\in\Lambda} x_\lambda \phi_\lambda$, with uniquely determined coefficients $x_\lambda \in \mathbb{R}$, the stated minimization is equivalent to

$$\begin{cases} \min\limits_{x} & \sum\limits_{\lambda\in\Lambda} |x_\lambda|^r \\ \text{s.t.} & \max\limits_{\lambda\in\Lambda} \dfrac{|\kappa_\lambda^{-1}\langle v_\lambda, g\rangle - x_\lambda|}{q_\lambda} \le 1 \, . \end{cases}$$

The solution of the latter optimization problem is given by component-wise soft-thresholding of $\kappa_\lambda^{-1}\langle v_\lambda, g\rangle$ with thresholds q_λ (see the proof of Theorem 3). This results in $\hat f^{(\text{WV})}$. □

B.2 Proofs in Section 4

In this subsection, we provide proofs for the assertions in Section 4 in the paper. They rely on the results from Appendix A.

To establish Theorem 8, we apply Theorem 18 to the case where R is the block ℓ^1-penalty. To that end we first compute its Fenchel conjugate. Recall that subsets Λ_a are assumed to be disjoint.

Lemma 19. *The Fenchel conjugate of the functional* $R(\beta) = \sum_{a\in\mathcal{A}} w_a \|\boldsymbol{\Phi}_a \beta\|_2$ *is given by*

$$R^*(\mu) = \begin{cases} 0 & \text{if } \max_{a\in\mathcal{A}} \frac{\|\boldsymbol{\Phi}_a \mu\|_2}{w_a} \le 1 \\ \infty & \text{otherwise}. \end{cases}$$

Proof. This can be verified by straightforward computation. □

We are now ready to prove Theorem 8.

Proof of Theorem 8. Theorem 18 implies that $\beta_\star \in \mathbb{R}^p$ is a minimizer of the adaptive group lasso in (19) if and only if it is a minimizer of the functional $\frac{1}{2}\|\boldsymbol{X}\beta\|_2^2 + R^*(\boldsymbol{X}^\top(Y - \boldsymbol{X}\beta))$ with the particular choice $R(\beta) = \gamma \sum_{a\in\mathcal{A}} w_a \|\boldsymbol{\Phi}_a \beta\|_2$. According to Lemma 19 this is equivalent to the fact that $\beta_\star$ minimizes $\frac{1}{2}\|\boldsymbol{X}\beta\|_2^2$ under the constraint

$$\max_{a\in\mathcal{A}} \frac{\|\boldsymbol{\Phi}_a \boldsymbol{X}^\top (Y - \boldsymbol{X}\beta)\|_2}{w_a} \le \gamma\,.$$

This means that $\beta_\star$ is the unique minimizer of (20) as we intended to show. □

Proof of Corollary 9. By specializing $(\phi_\lambda)_{\lambda\in\Lambda}$ to the standard basis in $\mathbb{R}^n$, letting the groups consist of single elements, and taking all weights equal to one, the constraint in Theorem 8 reduces to $\|\boldsymbol{X}^\top(Y - \boldsymbol{X}\beta)\|_\infty \le \gamma$. Further, in such a situation the adaptive group lasso reduces to the standard lasso. Hence the claim follows from Theorem 8. □

Proof of Corollary 10. According to Theorem 5, the estimators in (15) and (20) coincide. According to Theorem 8, the estimators in (20) and (19) coincide. □

Proof of Theorem 11. It follows immediately from Theorems 14 and 15. □

References

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.*, 22(4):351–361.

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1):115–129.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, 96(455):939–967. With discussion and a rejoinder by the authors.

Aspelmeier, T., Egner, A., and Munk, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.*, 2(1):163–202.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202.

Becker, S., Bobin, J., and Candès, E. J. (2011). NESTA: a fast and accurate first-order method for sparse recovery. *SIAM J. Imaging Sci.*, 4(1):1–39.

Behr, M., Holmes, C., and Munk, A. (2018). Multiscale blind source separation. *Ann. Statist.*, 46(2):711–744.

Bertero, M., Boccacci, P., Desiderà, G., and Vicidomini, G. (2009). Image deblurring with Poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, 26.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.*, 87(419):738–754.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924.

Cai, T. and Zhou, H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.*, 37(2):569–595.

Candès, E. J. and Donoho, D. (2002). Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Ann. Statist.*, 30(3):784–842.

Candès, E. J. and Donoho, D. L. (1999). Ridgelets: a key to higher-dimensional intermittency? *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.*, 357(1760):2495–2509.

Candès, E. J. and Donoho, D. L. (2000). Curvelets–a surprisingly effective nonadaptive representation for objects with edges. In Cohen, A., Rabut, C., and Schumaker, L. L., editors, *Curves and Surfaces*, pages 105–120. Vanderbilt University Press, Nashville, TN.

Candès, E. J. and Guo, F. (2002). New multiscale transforms, minimum total variation synthesis: applications to edge preserving image reconstruction. *Signal Process.*, 82:1519–1543.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*, volume 105 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition. A modern perspective.

Chambolle, A., DeVore, R. A., Lee, N., and Lucier, B. J. (1998). Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.*, 7(3):319–335.

Chambolle, A. and Lions, P.-L. (1997). Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188.

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145.

Chan, T. F. and Zhou, H. (2000). Total variation improved wavelet thresholding in image compression. In *IEEE Int'l Conf. on Image Process.*, volume 2, pages 391–394.

Chesneau, C., Fadili, J., and Starck, J.-L.-L. (2010). Stein block thresholding for wavelet-based image deconvolution. *Electron. J. Stat.*, 4:415–435.

Chui, C. K. (1992). *An introduction to wavelets*, volume 1 of *Wavelet Analysis and its Applications*. Academic Press, Inc., Boston, MA.

Cohen, A. (2003). *Numerical Analysis of Wavelet Methods*, volume 32 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam.

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant de-noising. In Antoniadis, A. and Oppenheim, G., editors, *Lecture Notes in Statistics*, volume 103, pages 125–150. Springer.

Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29(1):1–65.

Davies, P. L., Kovac, A., and Meise, M. (2009). Nonparametric regression, confidence regions and regularization. *Ann. Statist.*, 37(5B):2597–2625.

Davies, P. L. and Meise, M. (2008). Approximating data with weighted smoothing splines. *J. Nonparametr. Stat.*, 20(3):207–228.

Davis, R. A. and Yau, C. Y. (2013). Consistency of minimum description length model selection for piecewise stationary time series models. *Electron. J. Stat.*, 7:381–411.

del Álamo, M., Li, H., and Munk, A. (2021). Frame-constrained total variation regularization for white noise regression. *Ann. Statist. (To appear) arXiv preprint arXiv:1807.02038.*

del Álamo, M., Li, H., Munk, A., and Werner, F. (2020). Variational multiscale nonparametric regression: algorithms and implementation. *Algorithms*, 13(11):Paper No. 296, 24.

del Álamo, M. and Munk, A. (2020). Total variation multiscale estimators for linear inverse problems. *Inf. Inference*, 9(4):961–986.

Dette, H., Eckle, T., and Vetter, M. (2020). Multiscale change point detection for dependent data. *Scand. J. Stat.*, 47(4):1243–1274.

Do, M. N. and Vetterli, M. (2005). The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans. Image Process.*, 14(12):2091–2106.

Dong, Y., Hintermüller, M., and Rincon-Camacho, M. M. (2011). Automated regularization parameter selection in multi-scale total variation models for image restoration. *J. Math. Imaging Vision*, 40(1):82–104.

Donoho, D. L. (1995a). De-noising by soft-thresholding. *IEEE Trans. Inf. Theory*, 41(3):613–627.

Donoho, D. L. (1995b). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152.

Durand, S. and Froment, J. (2001). Artifact free signal denoising with wavelets. *Proceedings of ICASSP 2001 (26th International Conference on Acoustics, Speech, and Signal Processing)*, 6:3685–3688.

Eggermont, P. P. B. and LaRiccia, V. N. (2009). *Maximum penalized likelihood estimation. Volume II.* Springer Series in Statistics. Springer, Dordrecht. Regression.

Epstein, C. L. (2008). *Introduction to the mathematics of medical imaging.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition.

Fang, B., Guntuboyina, A., and Sen, B. (2021). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *Ann. Statist.*, 49(2):–.

Fang, X., Li, J., and Siegmund, D. (2020). Segmentation and estimation of change-point models: false positive control and confidence regions. *Ann. Statist.*, 48(3):1615–1647.

Frick, K., Marnitz, P., and Munk, A. (2012). Statistical multiresolution estimation in imaging: Fundamental concepts and algorithmic approach. *Electron. J. Statist.*, 6:231–268.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580. With 32 discussions by 47 authors and a rejoinder by the authors.

Fuchs, J. J. (2001). On the application of the global matched filter to doa estimation with uniform circular arrays. *IEEE Trans. Signal Process.*, 49(4):702–709.

Fuchs, J. J. (2004). On sparse representations in arbitrary redundant basis. *IEEE Trans. Inf. Theory*, 50(6):1341–1344.

Gao, H. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.*, 7(4):469–488.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Grasmair, M., Li, H., and Munk, A. (2018). Variational multiscale nonparametric regression: smooth functions. *Ann. Inst. Henri Poincaré Probab. Stat.*, 54(2):1058–1097.

Guntuboyina, A., Lieu, D., Chatterjee, S., and Sen, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.*, 48(1):205–229.

Hall, P., Penev, S., Kerkyacharian, G., and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Stat. Comput.*, 7:115–124.

Haltmeier, M. and Munk, A. (2014). Extreme value analysis of empirical frame coefficients and implications for denoising by soft-thresholding. *Appl. Comput. Harmon. Anal.*, 36(3):434–460.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer Series in Statistics. Springer-Verlag, New York.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica*, 18(4):1603–1618.

Hütter, J.-C. and Rigollet, P. (2016). Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146.

Ivanov, V. K., Vasin, V. V., and Tanana, V. P. (2002). *Theory of linear ill-posed problems and its applications*. Inverse and Ill-posed Problems Series. VSP, Utrecht, second edition. Translated and revised from the 1978 Russian original.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 361–379. Univ. California Press, Berkeley, Calif.

Kabluchko, Z. (2011). Extremes of the standardized Gaussian noise. *Stochastic Process. Appl.*, 121(3):515–533.

Kalifa, J. and Mallat, S. (2003). Thresholding estimators for linear inverse problems and deconvolutions. *Ann. Statist.*, 31(1):58–109.

Kerkyacharian, G., Kyriazis, G., Le Pennec, E., Petrushev, P., and Picard, D. (2010). Inversion of noisy Radon transform by SVD based needlets. *Appl. Comput. Harmon. Anal.*, 28(1):24–45.

König, C., Munk, A., and Werner, F. (2020). Multidimensional multiscale scanning in exponential families: limit theory and statistical consequences. *Ann. Statist.*, 48(2):655–678.

Korostelev, A. and Korosteleva, O. (2011). *Mathematical statistics*, volume 119 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI. Asymptotic minimax theory.

Kovács, S., Li, H., Haubner, L., Munk, A., and Bühlmann, P. (2020). Optimistic search strategy: change point detection for large-scale data via adaptive logarithmic queries. *arXiv preprint arXiv:2010.10194*.

Kutyniok, G., Shahram, M., and Zhuang, X. (2012). ShearLab: a rational design of a digital parabolic scaling algorithm. *SIAM J. Imaging Sci.*, 5(4):1291–1332.

Labate, D., Lim, W.-Q., Kutyniok, G., and Weiss, G. (2005). Sparse multidimensional representation using shearlets. *Proc. SPIE*, 5914:59140U.

Lang, M., Guo, H., Odegard, J. E., Burrus, C. S., and Wells, R. O., J. (1996). Noise reduction using an undecimated discrete wavelet transform. *IEEE Signal Process. Lett.*, 3(1):10–12.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, 34(5):2554–2591.

Lenzen, F. and Berger, J. (2015). Solution-driven adaptive total variation regularization. In *Scale space and variational methods in computer vision*, volume 9087 of *Lecture Notes in Comput. Sci.*, pages 203–215. Springer, Cham.

Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.

Li, H., Guo, Q., and Munk, A. (2019). Multiscale change-point segmentation: beyond step functions. *Electron. J. Stat.*, 13(2):3254–3296.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.*, 10(1):918–959.

Li, H., Munk, A., Sieling, H., and Walther, G. (2020). The essential histogram. *Biometrika*, 107(2):347–364.

Li, H. and Werner, F. (2020). Empirical risk minimization as parameter choice rule for general linear regularization methods. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(1):405–427.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.*, 27(2):415–438.

Ma, J. and Plonka, G. (2010). The curvelet transform. *IEEE Signal Process. Mag.*, 27(2):118–133.

Malgouyres, F. (2002a). Mathematical analysis of a model which combines total variation and wavelet for image restoration. *J. Inform. Process.*, 2(1):1–10.

Malgouyres, F. (2002b). Minimizing the total variation under a general convex constraint for image restoration. *IEEE Trans. Image Process.*, 11(12):1450–1456.

Mallat, S. (2009). *A wavelet tour of signal processing*. Elsevier/Academic Press, Amsterdam, third edition. The sparse way, With contributions from Gabriel Peyré.

Mallows, C. L. (2000). Some comments on c_p. *Technometrics*, 42(1):87–94.

Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423.

Morozov, V. A. (1966). On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In Antoniadis, A. and Oppenheim, G., editors, *Lecture Notes in Statistics*, volume 103, pages 281–299. Springer.

Natterer, F. and Wübbeling, F. (2001). *Mathematical methods in image reconstruction*. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Nemirovskii, A. S. (1985). Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet*, 3:50–60.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152.

Ortelli, F. and van de Geer, S. (2020). Adaptive rates for total variation image denoising. *J. Mach. Learn. Res.*, 21:247–238.

Ortelli, F. and van de Geer, S. (2021). Tensor denoising with trend filtering. *arXiv preprint arXiv:2101.10692*.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527. With comments and a rejoinder by the author.

Pein, F., Sieling, H., and Munk, A. (2017). Heterogeneous change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(4):1207–1227.

Pesquet, J.-C., Krim, H., and Carfantan, H. (1996). Time-invariant orthonormal wavelet representations. *IEEE Trans. Signal Process.*, 44(8):1964–1970.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, 16(1):356–366.

Proksch, K., Werner, F., and Munk, A. (2018). Multiscale scanning in inverse problems. *Ann. Statist.*, 46(6B):3569–3602.

Rockafellar, R. T. (1970). *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1–4):259–268.

Sadhanala, V., Wang, Y.-X., Sharpnack, J. L., and Tibshirani, R. J. (2017). Higher-order total variation classes on grids: Minimax theory and trend filtering methods. *Proc. NeurIPS*, 30.

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.*, 41(3):1299–1328.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434):655–665.

Sharpnack, J. and Arias-Castro, E. (2016). Exact asymptotics for the scan statistic and fast alternatives. *Electron. J. Stat.*, 10(2):2641–2684.

Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.*, 23(1):255–271.

Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.*, 37(3):1405–1436.

Starck, J.-L., Candès, E. J., and Donoho, D. L. (2002). The curvelet transform for image denoising. *IEEE Trans. Image Process.*, 11(6):670–684.

Starck, J.-L., Donoho, D., and Candès, E. (2011). Very high quality image restoration by combining wavelets and curvelets. *Proc. SPIE*, 4478:9.

Starck, J.-L., Murtagh, F., and Fadili, J. M. (2010). *Sparse image and signal processing*. Cambridge University Press, Cambridge.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.*, 42(1):285–323.

Tropp, J. A. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory*, 52(3):1030–1051.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Vanegas, L. J., Behr, M., and Munk, A. (2021). Multiscale quantile segmentation. *J. Amer. Statist. Assoc.*, 0(0):1–14.

Vasin, V. V. (1970). The connection between certain variational methods for the approximate solution of ill posed problems. *Mat. Zametki*, 7:265–272.

Vidakovic, B. (1999). *Statistical modeling by wavelets.* Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4):651–667.

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wainwright, M. J. (2019). *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge. A non-asymptotic viewpoint.

Walter, G. G. and Shen, X. (2001). *Wavelets and other orthogonal systems.* Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Walther, G. and Perry, A. (2020). Calibrating the scan statistic: finite sample performance vs. asymptotics. *arXiv preprint arXiv:2008.06136.*

Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.*, 17:Paper No. 105, 41.

Wendland, H. (2005). *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics.* Cambridge University Press, Cambridge.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 309.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.